Universidad
Carlos III de Madrid
www.uc3m.es

TESIS DOCTORAL

Smooth Generalized Linear Models for Aggregated Data

Autor:

Diego Ayma Anza

Director/es:

María Durbán

Dae-Jin Lee

DEPARTAMENTO DE ESTADÍSTICA

Universidad Carlos III de Madrid
www.uc3m.es

TESIS DOCTORAL

Smooth Generalized Linear Models for Aggregated Data

Autor: *Diego Ayma Anza*

Director/es: María Durbán y Dae-Jin Lee

Firma del Tribunal Calificador:

Firma

Presidente: Miguel Ángel Martínez Beneito

Vocal: Jutta Gampe

Secretario: Irene Albarrán Lozano

Calificación:

Leganés, de de

Universidad Carlos III de Madrid
Ph.D. Program in Mathematical Engineering

Ph.D. Thesis

Smooth Generalized Linear Models for Aggregated Data

Author:
Diego Ayma Anza

Advisors:
María Durbán
Dae-Jin Lee

Department of Statistics
Leganés, Madrid
October 2016

To my family,
for their endless support

Acknowledgements

I would like to thank my advisors María Durbán and Dae-Jin Lee, for their permanent support, encouragement, and guidance throughout my Ph.D. studies. Thanks for share your expertise, knowledge, and passion for statistical modelling. Without you, this thesis could probably never have come into being.

I also want to extend my gratitude to Paul H. C. Eilers, for his comments and suggestions on my work. Thanks for your kindness in sharing your knowledge about efficient programming.

I am grateful to the Department of Statistics of Universidad Carlos III de Madrid, for providing the financial support to carry out this thesis. In particular, I want to show my gratitude to Gema García, for her sympathy and administrative assistance, and Carmen Aguilera, Ignacio Cascos, and Elisa Molanes, for being nice people and sharing their teaching experience.

I want to express my sincere gratitude to Carlo G. Camarda and Nadine Ouellette, for their kindness and hospitality during my research stay at INED, and Anne Simon for let me stay in her cosy flat at Vincennes. I wish to thank Viorela Diaconu, Jenny García, and Silvia Méndez, for sharing many dinners, city walks, and laughs in Paris: *Ce fut un plaisir de vous rencontrer!*

I would like to thank all my Ph.D. fellows, particularly to the following people for their friendship and making my time in Madrid so much fun: Xiaolin Ayón, Alba Carballo, Lino (Gustavo) Garza, María Guadarrama, Ginette Lafit, Willian Oliveira, and Willy Ugaz. Also, I want to extend my gratitude to my Chilean fellows Katherine Gálvez, Javier González, Nathalie Humeniy, Pamela Pizarro, and Claudia Soto: *les deseo lo mejor a todos ustedes!*

Finally, and most importantly, I wish to thank my parents, Jorge Ayma and

Eliana Anza, and my brothers, Jorge Ayma and Benjamín Ayma, for their love and constant support throughout every step of my life. No matter what the future holds or the distance between us, I will always be for you ☺.

The work presented in this thesis was supported by the Spanish Ministry of Economy and Competitiveness grants MTM2011-28285-C02-02 and MTM2014-52184-P.

Abstract

Aggregated data commonly appear in areas such as epidemiology, demography, and public health. Generally, the aggregation process is done to protect the privacy of patients, to facilitate compact presentation, or to make it comparable with other coarser datasets. However, this process may hinder the visualization of the underlying distribution that follows the data. Also, it prohibit the direct analysis of relationships between aggregated data and potential risk factors, which are commonly measured at a finer resolution. Therefore, it is of interest to develop statistical methodologies that deal with the disaggregation of coarse health data at a finer scale. For example, in the spatial setting, it could be desirable to obtain estimates, from coarse areal data, at a fine spatial grid or units less coarser than the original ones. These two cases are known as the area-to-point (ATP) and area-to-area (ATA) cases, respectively, which are illustrated in the first chapter of this thesis. Moreover, we can have spatial data recorded at coarse units over time. In some cases, the temporal dimension can also be in an aggregated form, hindering the visualization of the evolution of the underlying process over time.

In this thesis we propose the use of a novel non-parametric method that we called *composite link mixed model* or, more succinctly, CLMM. In our proposed model, we look at the observed data as indirect observations of an underlying process (defined at a finer resolution than observed data), which we want to estimate. The mixed model formulation of our proposal allow us to include fine-scale population information and complex structures as random effects as parts of the modelling of the underlying trend. Since the CLMM is based on the approach given by Eilers (2007), called penalized composite link model (PCLM), we briefly review the PCLM approach in the first section of the second chapter of this thesis. Then, in the second section of this chapter, we introduce the CLMM approach under an univariate setting, which can be seen as a reformulation of the PCLM

into a mixed model framework. This is achieved by following the mixed model reformulation of P-splines proposed in Currie and Durbán (2002) and Currie et al. (2006), which is also reviewed here. Then, the parameter estimation of the CLMM can be done under the framework of mixed model theory. This offers another alternative for the estimation of the PCLM, avoiding the use of information criteria for smoothing parameter selection. In the third section of the second chapter, we extend the CLMM approach to the multidimensional (array) case, where Kronecker products are involved in the extended model formulation. Illustrations for the univariate and the multidimensional array settings are presented throughout the second chapter, using mortality and fertility datasets.

In the third chapter, we present a new methodology for the analysis of spatially aggregated data, by extending the CLMM approach developed in the second chapter to the spatial case. The spatial CLMM provides smoothed solutions for the ATP and ATA cases described in the first chapter, i.e., it gives a smoothed estimation for the underlying spatial trend, from aggregated data, at a finer resolution. The ATP and ATA cases are illustrated using several mortality (or morbidity) datasets, and simulation studies of the prediction performance between our approach and the area-to-point Poisson kriging of Goovaerts (2006) are realized. Also, in the third chapter we provide a methodology to deal with the overdispersion problem, which is based on the PRIDE ('penalized regression with individual deviance effects') approach of Perperoglou and Eilers (2010).

In the fourth chapter, we generalize the methodology developed in the third chapter for the analysis of spatio-temporally aggregated data. Under this framework, we adapt the SAP ('separation of anisotropic penalties') algorithm of Rodríguez-Álvarez et al. (2015) and the GLAM ('generalized linear array model') algorithms given in Currie et al. (2006) and Eilers et al. (2006), to the CLMM context. The use of these efficient algorithms allow us to avoid possible storage problems and to speed up the computational time of the model estimation. We illustrate the methodology presented in this chapter by using a Q fever incidence dataset recorded in the Netherlands at municipality level and by months. Our aim, then, is to estimate smoothed incidences at a fine spatial grid over the study area throughout the 53 weeks of 2009. A simulation study is provided at the end of chapter four, in order to evaluate the prediction performance of our approach under three different coarse situations, using a detailed (and confidential) Q fever

incidence dataset.

Finally, the fifth chapter summarizes the main contributions made in this thesis and further work.

Resumen

Datos agregados aparecen comúnmente en áreas como la epidemiología, demografía, y salud pública. Generalmente, el proceso de agregación es efectuado para proteger la privacidad de los pacientes, para facilitar una presentación compacta, o para hacerlos comparables con otros conjuntos de datos más gruesos. Sin embargo, este proceso puede dificultar la visualización de la distribución subyacente que siguen los datos. Además, prohibe el análisis directo de relaciones entre los datos agregados y factores de riesgos potenciales, los cuales son medidos usualmente en una resolución más fina. En consecuencia, es de interés el desarrollar metodologías estadísticas que traten la desagregación de datos de salud gruesos a una escala más fina. Por ejemplo, en el caso espacial, podría ser deseable obtener estimaciones, a partir de datos disponibles en unidades geográficas gruesas, en una malla espacial fina o en unidades menos gruesas que las originales. Estos dos casos se conocen como los casos área-a-punto (ATP, 'area-to-point') y área-a-área (ATA, 'area-to-area'), respectivamente, los cuales son ilustrados en el primer capítulo de esta tesis. Mas aún, podemos tener datos espaciales registrados en unidades geográficas gruesas a lo largo del tiempo. En algunos casos, la dimensión temporal también puede estar en una forma agregada, dificultando la visualización de la evolución del proceso subyecente a lo largo del tiempo.

En esta tesis proponemos el uso de un novedoso método no-paramétrico que llamamos modelo mixto de enlace compuesto o, más brevemente, CLMM ('composite link mixed model'). En nuestro modelo propuesto, miramos a los datos observados como observaciones indirectas de un proceso subyacente (definido en una resolución más fina que los datos observados), el cual queremos estimar. La formulación de modelo mixto en nuestra propuesta nos permite incluir información de la población medida en una escala fina y estructuras complejas como efectos aleatorios, como partes de la modelización de la tendencia subyacente. Dado que

el CLMM está basado en el enfoque dado por Eilers (2007), llamado modelo de enlace compuesto penalizado (PCLM, 'penalized composite link model'), revisaremos brevemente el enfoque PCLM en la primera sección del segundo capítulo de esta tesis. Luego, en la segunda sección de este capítulo, introduciremos el enfoque CLMM bajo un marco univariante, el cual puede ser visto como una reformulación del PCLM en un marco de modelo mixto. Esto es logrado siguiendo la reformulación como modelo mixto de los P-splines propuestos por Currie y Durbán (2002) y Currie et al. (2006), el cual es también revisado aquí. Luego, la estimación de parámetros del CLMM puede hacerse bajo el marco de la teoría de los modelos mixtos. Esto ofrece otra alternativa para la estimación del PCLM, evitando el uso de criterios de información para la selección del parámetro de suavizado. En la tercera sección del segundo capítulo, extendemos el enfoque CLMM al caso (array) multidimensional, en donde productos de Kronecker están implicados en la formulación del modelo extendido. Ilustraciones para los casos univariantes y (array) multidimensional son presentados a lo largo del segundo capítulo, usando conjuntos de datos de mortalidad y fertilidad.

En el tercer capítulo, presentamos una nueva metodología para el análisis de datos agregados espacialmente, extendiendo el enfoque CLMM desarrollado en el segundo capítulo al caso espacial. El CLMM espacial proporciona soluciones suavizadas para los casos ATP y ATA descritos en el primer capítulo, es decir, entrega una estimación suavizada para la tendencia espacial subyacente, a partir de datos agregados, en una resolución más fina. Los casos ATP y ATA son ilustrados usando diferentes conjuntos de datos de mortalidad (o morbilidad), y estudios de simulación sobre el desempeño de predicción entre nuestro enfoque y el Poisson kriging área-a-punto de Goovaerts (2006) son realizados. Además, en el tercer capítulo proporcionamos una metodología para lidiar con el problema de sobredispersión, el cual está basado en el enfoque PRIDE ('penalized regression with individual deviance effects') de Perperoglou y Eilers (2010).

En el cuarto capítulo, generalizamos la metodología desarrollada en el tercer capítulo para el análisis de datos agregados espacio-temporalmente. Bajo este contexto, adaptamos el algoritmo SAP ('separation of anisotropic penalties') de Rodríguez-Álvarez et al. (2015) y los algoritmos GLAM ('generalized linear array model') dados por Currie et al. (2006) y Eilers et al. (2006) en el contexto de los CLMMs. El uso de estos algoritmos eficientes nos permite evitar posibles

problemas de almacenamiento y acelerar el tiempo de cómputo de la estimación del modelo. Ilustramos la metodología presentada en este capítulo usando un conjunto de datos sobre incidencia de fiebre Q registradas en Holanda a nivel municipal y por meses. Nuestro objetivo, luego, es el de estimar incidencias suavizadas en una malla espacial fina sobre el área de estudio a lo largo de las 53 semanas del 2009. Un estudio de simulación es dado al final del cuarto capítulo, de manera de evaluar el desempeño de predicción de nuestro enfoque bajo tres diferentes situaciones de agregación, usando un conjunto de datos detallado (y confidencial) de incidencia de fiebre Q.

Finalmente, el quinto capítulo resume las contribuciones principales hechas en esta tesis y el trabajo a futuro.

Contents

List of figures

Chapter 1

Introduction

1.1 Background

Spatial or temporal data are often collected at several scales. Time series of counts, histograms, data from satellites, or disease registries are common examples of data that are recorded at different layers and, in many occasions, they are *incompatible*. For example, the levels of a certain contaminant may be recorded at several monitoring stations, while the number of people affected by exposure is provided at post code level to preserve privacy. Another situation arises when working with historical death records: they often use wide intervals that narrow down over time. In this case, the aim might be to estimate a smooth mortality distribution using population records that are often more precise. We could summarize these situations under the *incompatible data problem*. In the case of spatial data, the aim might be to estimate the distribution of the outcome at a new level of spatial aggregation. If we are dealing with areal or regional data (i.e., data recorded over irregular geographical units, like counties, districts, and municipalities), this is called the *modifiable areal unit problem* (MAUP) and in the case of data modelled through a spatial process, it is called the *change of support problem* (COSP).

Areal data frequently appear in areas such as epidemiology, demography, and public health. Methodological contributions for the analysis of these type of data have been benefited by the advances in geographic information systems (GISs), the access to reliable health data registers, and the disposition of powerful software capable to processing and analysing large amount of data.

Among the types of epidemiological enquiries, disease mapping has received a great interest in public health, since allows the visualization of the spatial distribution that a mortality (or morbidity) risk of a disease has in a specific study area. This is carried out by means of *disease maps*, which are not only used for descriptive purposes, but also for surveillance to highlight areas of excess, and to aid policy formation and resource allocation (Elliott and Wartenberg, 2004).

In general, rates are used as measures of the risk, since they incorporate information about the population of each geographical unit. A first attempt to estimate the relative risk within each unit is the so-called *standardized mortality* (or *morbidity*) *rate* (SMR). The SMR for each unit v_i is calculated as:

$$\text{SMR}_i = \frac{y_i}{e_i}, \text{ for } i = 1, ..., n, \tag{1.1}$$

where y_i and e_i are the observed and expected number of deaths (or incidents cases of disease) at unit v_i, respectively, and n is the total number of units that form the study area. The SMRs given in Eq. (1.1) can also be provided on a logarithmic scale (i.e., log(SMRs)), or as percentages (i.e., multiplied by 10^2). In order to depict the spatial distribution of SMRs (or any areal data), a *choropleth map* is commonly used. This type of disease map uses a color palette to depict different values of the attribute variable (the SMR in this case) associated with each unit. Thus, each unit is colored according to the class into which its corresponding attribute variable falls (Waller and Gotway, 2004).

Figure 1.1 shows a choropleth map of log(SMRs) by sudden infant death syndrome (SIDS) in North Carolina, USA, during the period 1974-1978, which are recorded at county level (100 counties in total). The SIDS dataset (Cressie, 1993) has been analysed by many researchers and incorporated in several statistical software packages, such as (R Core Team, 2015). See Bivand et al. (2008) and the package (Bivand and Lewin-Koh, 2016) for more information about this dataset.

For the legend in Figure 1.1, we selected the deciles of the raw log(SMRs) as the class boundaries. The colors used for the legend classes are based on the palette developed by Brewer et al. (2003), which is available in the package (Neuwirth, 2014). Thus, higher log(SMRs) than the median tend to be more dark, while lower log(SMRs) tend to be more clear.

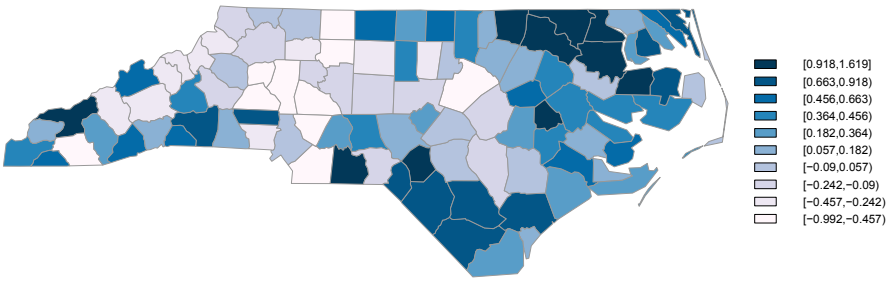


Figure 1.1: Raw log(SMRs) for the North Carolina SIDS dataset (period 1974-1978).

However, choropleth maps (such as in Figure 1.1) must be interpreted with caution: ratios calculated from small or sparsely populated units are likely to be elevated artificially (Waller and Gotway, 2004). This effect, known as the *small number problem*, may hinder the detection of meaningful patterns in the study area. Another problem that can arise is the *spatial misalignment* between potential risk factors and health data: in general, the former are available on a finer spatial resolution than the latter. For example, most deprivation indices are built on the smallest possible geographical units of a certain region (see Rey et al., 2009; Salmond and Crampton, 2012) or even on a fine grid (Caudeville et al., 2012). Environmental agents (such as air pollution) constitute examples of potential risk factors that vary continuously in space. Consequently, this issue precludes their direct use in a correlation analysis, which is a critical step for disease control intervention. Therefore, it is relevant to develop spatial methodologies that filter the noise caused by the small number problem and allow the creation of mortality maps, from aggregated data, at a finer spatial resolution.

Different approaches have been used to reduce the noise in spatially aggregated mortality rates (see Besag et al., 1991; MacNab and Dean, 2002; Fahrmeir et al., 2004; Goovaerts, 2005; Lee and Durbán, 2009; among others). However, they give smoothed mortality estimates that are assumed constant over each unit, yielding a coarse spatial trend. To obtain a more detailed insight of mortality through units, several methodologies have been proposed in the literature. In a geostatistical framework, Kelsall and Wakefield (2002) obtained pointwise posterior medians of the underlying continuous risk surface, for colorectal cancer mortality in the UK

district of Birmingham, via a Gaussian random field (GRF) model. Goovaerts (2006) generalized the Poisson kriging algorithm given by Monestiez et al. (2005, 2006), which incorporates the size and shape of the units, as well as the population density, into the filtering of noisy mortality rates. This generalization allows the mapping of the corresponding mortality risk at a fine resolution. The performance of his approach, called area-to-point Poisson kriging, was compared with two geostatistical methods. The first one corresponds to the simple interpolation of raw rates to the nodes of a fine grid using ordinary kriging. The second one corresponds to the approach proposed by Berke (2004), in which the raw rates are replaced by their global empirical Bayes estimates before the interpolation process. Local Bayes estimates were also considered in the analysis, to attenuate the smoothing effect produced by the global mean term in the calculation of those Bayes estimates. Lately, and from a Bayesian inferential viewpoint, Diggle et al. (2013) used the class of log-Gaussian Cox processes (as models for spatial point process data) to construct a continuous map of lung cancer mortality risks in the Castile-La Mancha region of Spain, from spatially discrete data.

On the other hand, the incorporation of the temporal dimension in disease mapping enables the study of mortality risk (or disease incidence) evolution in each unit, during a certain period of time (generally divided in years). In this case, a *dynamic disease map* is used to depict such evolution. However, its inclusion implies a challenge when it comes to smoothing data, in terms of computational time and storage. Several techniques has been proposed for the spatio-temporal smoothing of health data; most of them developed under an empirical Bayes approach where B-splines are used (MacNab and Dean, 2001; Ugarte et al., 2010) or a hierarchical Bayesian framework where conditional autoregressive (CAR) structures are included (Waller et al., 1997; Martínez-Beneito et al., 2008). Within the latter approach, methods using integrated nested Laplace approximations (INLA, Rue et al., 2009) has recently been proposed (see Schrödle and Held, 2011; Ugarte et al., 2014; Bauer et al., 2016; among others).

All the works cited above provide smoothed estimates that are assumed constant over each unit and year. Also, most of them can be extended in order to include explanatory variables, which must be at the same spatio-temporal resolution as health data. Thus, they restrict the direct incorporation of fine-scale population information and other relevant risk factors recorded at a finer reso-

lution. Therefore, it is important to develop spatio-temporal methodologies that allow the creation of dynamic mortality maps, from spatio-temporally aggregated data, at a desirable fine resolution. As far as we know, there are no methodologies addressing the problem of disaggregation of health data both in space and time (although there exist works about the spatio-temporal disaggregation of Gaussian data; see, for example, Prairie et al., 2007; Segond et al., 2007; Schleiss and Berne, 2012; Bindhu and Narasimhan, 2015).

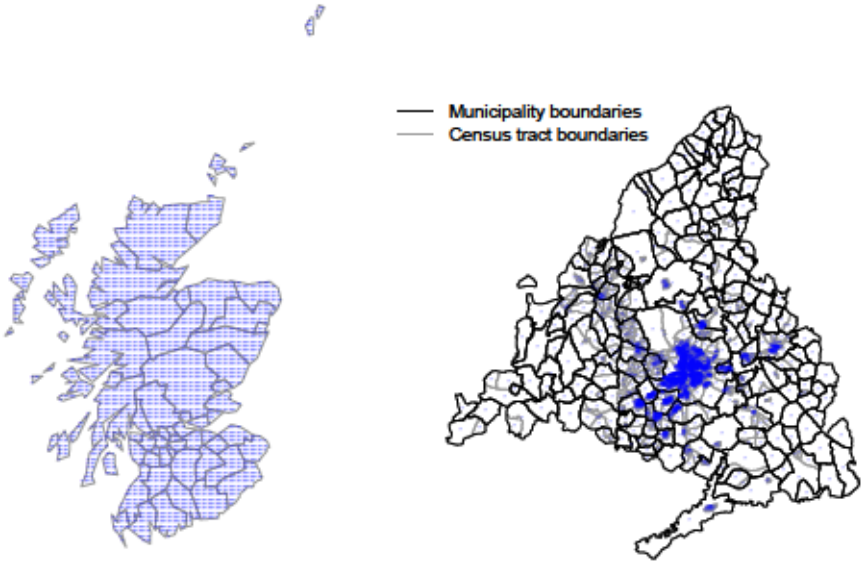
1.2 The smoothing approach

In this thesis we propose the use of a novel methodology that we called the composite link mixed model (CLMM) approach. The CLMM allows us to create mortality risk or disease incidence maps, from aggregated health data, at a desirable fine resolution, and to incorporate fine-scale information into the filtering of noisy rates. Under the CLMM approach, we look at the observed outcomes as observations of a latent or underlying process (i.e., as *indirect observations*) that we want to estimate. Also, we assume that the latent process behind aggregated data is smooth.

The flexibility of our approach is provided by the use of B-splines, together with a discrete penalty on the regression coefficients, following the P-spline methodology given by Eilers and Marx (1996). The mixed model structure in our proposal makes it possible to include specific random effects or further correlation structure if necessary, and to estimate the parameters of the CLMM under the framework of mixed model theory.

In a spatial context, we can have two types of disaggregation: 1) from coarse geographical units to a fine grid, i.e., the area-to-point (ATP) case, and 2) from coarse geographical units to smaller ones, i.e., the area-to-area (ATA) case. The CLMM approach can handle both cases in a nice way, by defining the spatial support of the underlying process in one way or another. These cases are illustrated in Figure 1.2 using two different maps, which we describe below.

Figure 1.2a illustrates the ATP case, where the coarse geographical units correspond to Scottish counties (56 counties in total). These counties define the spatial support for the well-known Scottish lip cancer dataset given by Clayton and Kaldor (1987). The goal, then, is to obtain a continuous surface (i.e., an *iso-*



(a) Area-to-point case (b) Area-to-area case

Figure 1.2: Illustrations of the area-to-point and area-to-area cases. The left map shows 56 Scottish counties in 1975, where a 120×120 fine grid was imposed. The 3885 blue points are the grid points that fall inside this map. The right map shows the municipalities (179) and census tracts (3906) of the Community of Madrid in 2001, where the blue points depict the centroid coordinates of the census tracts.

pleth map[1]) without subjective geographical boundaries (county boundaries in this case). For that purpose, a fine spatial grid (imposed over the map) is used as the fine spatial resolution for the underlying process. In Figure 1.2a, we have imposed a 120×120 fine grid over the map, where 3885 grid points (depicted in blue) fall inside the counties. The advantage of producing isopleth maps is to reduce the visual bias associated with the interpretation of choropleth maps (Cressie, 1993), which is produced by the variation in shape and size of units.

Figure 1.2b illustrates the ATA case, where the coarse units correspond to 179 municipalities in the Community of Madrid (CM). Here, the goal is to estimate a more refined spatial trend at census tract level (3906 census tracts in total). For that, the geographical centroid coordinates (depicted in blue) are used as

[1]Isopleth maps differ from choropleth maps in that the data are not grouped to a predefined region like a state or country. Temperature, for example, works better as an isopleth map than a choropleth map, because temperature is continuous but does not change abruptly at any point.

the spatial support for the underlying process at census tract level. Notice that most of the census tracts in Figure 1.2b are concentrated in the center of the CM, specifically in the municipality of Madrid. Indeed, this municipality in 2001 had 2358 census tracts. The resulting ATA estimates are, then, displayed in a choropleth map, which will offer a detailed insight of the process behind aggregated health data at census tract level.

We have to pointed out that the ATP and ATA cases are specific situations of the COSP, since in both cases we seek to obtain mortality risk estimates at a fine resolution from data available at coarse geographical units. Several solutions for the general COSP have been proposed depending on the following classification (see Gotway and Young, 2002, for a detailed description):

1) Point-to-point: Observe point data $Y(s_i)$ at location s_i, $i = 1, ..., n$, and the interest is about the process at new locations s_j^*, $j = 1, ..., m$. This has been addressed by spatial kriging (Cressie, 1993, Ch. 3) or Hierarchical geostatistical models (see Wikle et al., 1998, or Banerjee et al., 2015, for a summary of these methods), or cokriging (Chilés and Delfiner, 1999).

2) Point-to-area: Observe point data $Y(s_i)$ at location s_i, $i = 1, ..., n$, and the interest is on the process at a collection of areal units, v_j, $j = 1, ..., m$. The methods proposed include the use of areal centroids, spatial smoothing (see, for example, Müller et al., 1997) or block kriging (Goovaerts, 1997).

3) Area-to-area: We have observations associated with areal units $Y(v_i)$, $i = 1, ..., n$, and we want to infer about observations in other areas $Y(v_j^*)$, $j = 1, ..., m$. This problem have several ramifications depending on whether the new areas v_j^* are nested within the v_i's (Mugglin and Carlin, 1998) or are a separate partition (Banerjee et al., 2015).

4) Area-to-point: We have observations associated with areal units $Y(v_i)$, $i = 1, ..., n$, and we seek to infer about the process at certain locations s_j^*, $j = 1, ..., m$. This is the hardest problem and only few solutions (compared with the other cases) have been proposed, most of them are variations of the area-to-point kriging introduced by Kyriakidis (2004), and later applied to the Poisson case by Goovaerts (2006).

In the case of spatial data, the P-spline approach has already being used for the point-to-point case (see, for example, Lee and Durbán, 2011) and in the area-to-area case, when the aim is to predict the outcome of new areas out of the original region (Opsomer et al., 2008) or forecast future values (Ugarte et al., 2010). In the point-to-area case, there have been some results when dealing with a point process by imposing a regular fine grid over a map, counting the number of observations in each cell, and smoothing over the grid (van der Hoek et al., 2010).

In a spatio-temporal context, we will shows how the CLMM approach can deal with the simultaneous disaggregation of health data in space and time. For example, if we have data aggregated by counties and months, the CLMM is capable to estimate spatio-temporal trends at a fine spatial grid along weeks. However, this leads to an increase of the computational burden (in the CLMM estimation) and problems with data storage. In this thesis we propose solutions for these problems by using efficient algorithms proposed in the literature, which are adapted to the CLMM setting. Then, the resulting CLMM estimates are displayed in a dynamic map, allowing to visualize the evolution of the underlying process behind aggregated data at a fine resolution.

The CLMM approach and the extensions presented in this thesis were implemented in the statistical software . This free software environment allows to describe and analyse public health data, using a battery of packages, and to add additional functionality (provided by the users) by defining new functions. Thus, it offers a flexible tool for the programming of new statistical methodologies.

1.3 Dissertation structure

This thesis is organized as follows. In Chapter 2 we introduce the composite link mixed model approach in a univariate setting. We present details about the estimation of the model, and useful extensions to the multidimensional case. This part of the thesis generalizes the approach given in Eilers (2007), called penalized composite link model, into the mixed model framework. In Chapter 3 we present the composite link mixed model approach for spatially aggregated health data. Here, we discuss how the our proposal can handle the area-to-point and area-to-area cases illustrated in Section 1.2. Also, we extend our methodology in order to deal with the problem of overdispersion, often present in count data, and we

compare our proposal with the area-to-point Poisson kriging of Goovaerts (2006). Part of the work presented in this chapter is already published in Ayma et al. (2016). In Chapter 4 we generalize the presented methodology to the spatio-temporal setting, where efficient algorithms are presented and adapted under our approach. Finally, in Chapter 5 we summarize the main contributions given in this thesis and suggest possible future work.

Chapter 2

Composite link mixed models

In this chapter we present the composite link mixed model, which can be seen as a generalization of the penalized composite link model (PCLM) introduced by Eilers (2007). In Section 2.1 we introduce the PCLM methodology for the univariate case, focusing specially on the modelling of indirect observations of counts. In Section 2.2 we present more details about the methodology and its reformulation as a mixed model. This section is the heart of this chapter, and will allow us to understand useful extensions given in the following chapters. Finally, Section 2.3 shows how we can extend the composite link mixed model to the multidimensional case. All the examples provided in this chapter are used in order to illustrate the presented methodology.

2.1 Penalized composite link models: an introduction

The PCLM approach of Eilers (2007) is based on the model proposed by Thompson and Baker (1981), called composite link model (CLM). The CLM offers an elegant way to estimate the underlying or latent process behind observed grouped data, which can be seen as *indirect observations* of that process. It extends the generalized linear model (GLM, Nelder and Wedderburn, 1972) by associating more than one linear predictor with each observation, using the so-called composite link functions. These type of functions have been used in forest growth modelling (Candy, 1989, 1997), missing/incomplete categorical data analysis (Rindskopf,

1992; Galecki et al., 2001), randomized response modelling (van den Hout et al., 2010), paired comparison data analysis with missing responses (Dittrich et al., 2012), and multilevel and latent variable modelling (Rabe-Hesketh and Skrondal, 2007).

The theory of GLMs is well established in the statistical community, offering a unified way to build and estimate models for many type of observations (see McCullagh and Nelder, 1989). Due its popularity, the GLM approach has been implemented in several software packages, such as R and MATLAB©, and is included as the starting point of many statistical modelling books (see, for example, Wood, 2006a; Faraway, 2006). On the contrary, the CLM has not been received the same attention. This was acknowledged by Eilers (2012), where the author pointed out that some types of CLMs can run into numerical problems due their ill-conditioned nature (that is, there is no enough information in the data to estimate their parameters reliably). To overcome this situation, Eilers (2007) proposes to impose smoothness on the solution, by penalizing the log-likelihood with a roughness measure. Recently, his proposal has lead to several works related with digit preference and misreporting probabilities (Camarda et al., 2008), latent density estimation from grouped continuous data (Lambert and Eilers, 2009) and its extension to the bivariate case (Lambert, 2011), estimation of survival functions and hazard ratios from interval-censored data (Yavuz and Lambert, 2011), haplotype probabilities estimation from observed genotypes (Uh and Eilers, 2011), removement of artifacts in X-ray diffraction scans (de Rooi et al., 2014), ungrouping binned data (Rizzi et al., 2015), decomposition of complex series of counts (Camarda et al., 2016), among others.

The PCLM can be seen as the combination of the CLM of Thompson and Baker (1981) and the P-spline methodology given by Eilers and Marx (1996). In the next subsections we describe these two ingredients in order to understand the idea behind the PCLM approach.

2.1.1 The composite link model framework

Let $\boldsymbol{y} = (y_1, ..., y_n)'$ be a vector of observed responses whose components are assumed independently and identically distributed (in the exponential family) with means $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)'$. In the context of GLMs, the mean of an observation and

its linear predictor, denoted as η_i, are related through the link function, g, as $g(\mu_i) = \eta_i$. However, in some cases we wish to associate more than one linear predictor η_j (i.e., more than one set of observed explanatory variables, x_{jk}, with $k = 1, ..., p$) with each observation y_i. For example, when we have a histogram in which counts are aggregated in intervals, the total count is the result of the contribution of several latent observations. In order to achieve this, we suppose that $\gamma_j = g^{-1}(\eta_j)$, $j = 1, ..., m$, where m is the number of latent observations, and that $\mu_i = c_i(\gamma)$, with $\gamma = (\gamma_1, ..., \gamma_m)'$, where the c_i are known functions called composite link functions. Assuming that μ_i is a linear combination of the elements of γ, we have that $\mu_i = \sum_{j=1}^{m} c_{ij}\gamma_j$. Then, the CLM is given as:

$$\mathbb{E}(\boldsymbol{y}) = \boldsymbol{\mu} = \mathbf{C}\boldsymbol{\gamma}, \tag{2.1}$$

where $\mathbf{C}$ is a matrix of dimension $n \times m$ with entries c_{ij} and $\gamma = g^{-1}(\boldsymbol{\eta})$, with $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\theta}$ and $\mathbf{X}$ denoting a matrix of dimension $m \times p$ with entries x_{jk}.

The matrix $\mathbf{C}$ in Eq. (2.1) is called *composition matrix* and describes how the elements of latent vector γ are combined to yield $\boldsymbol{\mu}$. Thus, its structure will depend on the underlying process that generates the observed data. As an example of the form that a composition matrix may have, we suppose a grouping of age classes in a table of death counts. In this case, γ represents the expected number of deaths per one-year age classes and $\boldsymbol{y}$ the available data as totals of five-year age classes. If γ covers the ages from 1 to 100, then the mean $\boldsymbol{\mu}$ of the grouped counts has length 20 and the composition matrix $\mathbf{C}$ has dimension 20×100:

$$\mathbf{C} = \begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & 0 & \ddots & 0 & 0 & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 1 & \cdots & 1 \end{bmatrix}. \tag{2.2}$$

In (2.2), the elements of $\mathbf{C}$ are zero, except for those $c_{ij} = 1$ that indicate the elements of γ that are aggregated to form μ_i, $i = 1, ..., 20$. This example is a special case of grouped counts related with histograms that have coarse bins. Others examples involving HIV back-calculation, Poisson mixture, and present status data, can be found in Eilers (2007).

Since we are interested in the CLM application to grouped count data, we show

the structure of this model under the Poisson context. Suppose that we observe a vector of aggregated counts $\boldsymbol{y}$ that follows a Poisson distribution with mean vector $\boldsymbol{\mu}$. We want to estimate the latent distribution $\boldsymbol{\gamma}$ behind this aggregated data. Then, the model given in Eq. (2.1) becomes:

$$\boldsymbol{\mu} = \mathbf{C}\boldsymbol{\gamma} = \mathbf{C}\exp(\boldsymbol{\eta}), \tag{2.3}$$

with $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\theta}$. Notice that the non-negativity of the values of $\boldsymbol{\gamma}$ is guaranteed due its definition. As is indicated by Eilers (2012), the model given in Eq. (2.3) can be quite different from a process where samples from Poisson distributions with expectations $\boldsymbol{\gamma}$ are generated and then linearly combined with the composition matrix $\mathbf{C}$. This is because, in the latter case, when rows of $\mathbf{C}$ overlap, we cannot assume independence of the elements of $\boldsymbol{y}$.

To estimate the Poisson CLM given in Eq. (2.3), we follow the method of maximum likelihood (ML), which is usually used in the theory of GLMs (see, for example, Pawitan, 2001). For that, we consider the probability density function associated to y_i:

$$f(y_i;\mu_i) = \frac{\mu_i^{y_i}\exp(-\mu_i)}{y_i!} = \exp\left(y_i\log(\mu_i) - \mu_i - \log\left(y_i!\right)\right).$$

Then, the log-likelihood function based on n observations $\boldsymbol{y}$ is given by:

$$\ell = \sum_{i=1}^{n}\log(f(y_i;\mu_i)) = \sum_{i=1}^{n}\left(y_i\log(\mu_i) - \mu_i - \log(y_i!)\right). \tag{2.4}$$

From Eq. (2.3), we see that the estimation of $\boldsymbol{\gamma}$ (and, consequently, $\boldsymbol{\mu}$) is determined by the vector of regression coefficients $\boldsymbol{\theta}$. Therefore, by deriving the log-likelihood given in Eq. (2.4) with respect to each element of $\boldsymbol{\theta}$, we obtain that:

$$\frac{\partial\ell}{\partial\theta_k} = \sum_{i=1}^{n}\frac{y_i-\mu_i}{\mu_i}\frac{\partial\mu_i}{\partial\theta_k},$$

for $k = 1, ..., p$. Since $\mu_i = \sum_{j=1}^{m} c_{ij}\gamma_j$, we have that:

$$\frac{\partial\mu_i}{\partial\theta_k} = \sum_{j=1}^{m} c_{ij}\frac{\partial\gamma_j}{\partial\theta_k} = \sum_{j=1}^{m} c_{ij}x_{jk}\gamma_j,$$

and, thus, the ML equations are given by:

$$\sum_{i=1}^{n}(y_i - \mu_i)\breve{x}_{ik} = 0,$$

for $k = 1, ..., p$, where $\breve{x}_{ik} = \sum_{j=1}^{m} c_{ij}x_{jk}\gamma_j/\mu_i$. We can rewrite the previous system of equations in matrix form as:

$$\breve{\mathbf{X}}'(\boldsymbol{y} - \boldsymbol{\mu}) = \mathbf{0}, \tag{2.5}$$

where $\breve{\mathbf{X}} = \mathbf{W}^{-1}\mathbf{C}\boldsymbol{\Gamma}\mathbf{X}$, with $\mathbf{W} = \text{diag}(\boldsymbol{\mu})$ and $\boldsymbol{\Gamma} = \text{diag}(\boldsymbol{\gamma})$. Notice that the previous system of equations is nonlinear in $\boldsymbol{\theta}$ and an iterative procedure is needed to solve them. The iteratively re-weighted least square (IRLWS) algorithm can be used in this context, which is also used for the estimation of GLMs (for more details, see McCullagh and Nelder, 1989). The resulting IRWLS equations (expressed in matrix form) that solve the system of equations in Eq. (2.5) is:

$$\breve{\mathbf{X}}'\tilde{\mathbf{W}}\breve{\mathbf{X}}\widehat{\boldsymbol{\theta}} = \breve{\mathbf{X}}'\tilde{\mathbf{W}}(\tilde{\mathbf{W}}^{-1}(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}) + \breve{\mathbf{X}}\tilde{\boldsymbol{\theta}}), \tag{2.6}$$

where $\tilde{\mathbf{W}} = \text{diag}(\tilde{\boldsymbol{\mu}})$. Here, and hereafter, a tilde as in $\tilde{\boldsymbol{\theta}}$ indicates the current approximation to the solution and $\widehat{\boldsymbol{\theta}}$ denotes the updated estimate of $\boldsymbol{\theta}$. Notice that the IRLWS equations in Eq. (2.6) have exactly the same structure as that for a GLM. The difference is that in a GLM we would have $\mathbf{X}$ while here $\breve{\mathbf{X}}$ appears. Thus $\breve{\mathbf{X}}$ can be interpreted as a 'working' matrix $\mathbf{X}$.

As is stated by Eilers (2007), the direct application of Eq. (2.6) generally will not work for ill-posed data: the equations will be singular, when $p > n$, or severely ill-conditioned otherwise. To overcome this problem, Eilers (2007) proposes the introduction of a roughness penalty on the vector of regression coefficients $\boldsymbol{\theta}$ and the use of a B-spline basis instead of $\mathbf{X}$. This approach is known as the P-spline methodology, which was developed by Eilers and Marx (1996). Within this new framework, the latent vector $\boldsymbol{\gamma}$ is smooth and can be interpreted as a continuous distribution. In the next subsection we briefly describe in which consists this methodology.

2.1.2 The P-spline methodology

There are several techniques for fitting a smooth function that relates a response variable with a single predictor (see, for example, Hastie and Tibshirani, 1990, or Simonoff, 1996, for a good summary of them). We can see these smoothing models as a generalization of the linear regression model, which allow to estimate the function more precisely, but eventually with an added computational cost. Among them, there is an important group of models that use splines, which are piecewise polynomials that join at certain points called knots. Moreover, there are two main families within this group: 1) regression splines, and 2) smoothing splines. In the first one, it is necessary to select the number and localization of knots (in order to control the smoothness of the fitted function) and impose restrictions such that the piecewise polynomials join smoothly. Then, once we have made that choice, the model is adjusted by least squares. In general, the main drawback of regression splines is that we need to use complex algorithms for knots selection and, as a consequence, it is difficult to extended them to the multidimensional case. The second one appears as the solution to the problem of finding the function (with two continuous derivatives) that minimizes the penalized sum of squares, where the penalty term is related to the second derivative of the smooth function (more details are given in Green and Silverman, 1994). A drawback of smoothing splines is that they use the same number of knots (and hence, parameters) as observations. Splines with penalties, or commonly called P-splines (Eilers and Marx, 1996), deal with those drawbacks: they are low rank smoothers, i.e, the number of knots used is much less than the dimension of the data (making them computationally more efficient), and the introduction of penalties relaxes the importance of the choice of the number and localization of knots. Also, this approach is preferred instead of other smoothing techniques since it can be extended to the GLM framework in a straightforward way and it lacks of unwanted boundary effects.

In the rest of this subsection we review the P-spline methodology that allows to introduce the PCLM approach in the next subsection.

P-splines for unidimensional data

Suppose that we observe data pairs (x_i, y_i), $i = 1, ..., n$. A smooth relationship between the response variable $\boldsymbol{y} = (y_1, ..., y_n)'$ and a single predictor $\boldsymbol{x} = (x_1, ..., x_n)'$

is given by:

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \backsim \mathcal{N}(\mathbf{0}, \sigma_{\epsilon}^2 \mathbf{I}), \tag{2.7}$$

where $f(\cdot)$ is a unknown function of $\boldsymbol{x}$ assumed to be smooth and the components of $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)'$ are independent and identical distributed errors with variance σ_{ϵ}^2. The aim is to estimate function $f(\cdot)$ given the observed pairs (x_i, y_i). As it was introduced previously, there are several techniques for fitting the smooth function $f(\cdot)$. Here we focus on P-splines developed by Eilers and Marx (1996). Their methodology can be summarized as follows: 1) use a B-spline regression basis, and 2) introduce difference penalties over adjacent regression coefficients (by modifying the likelihood function), to control the smoothness of the fit.

In the model given in Eq. (2.7), we consider:

$$f(\boldsymbol{x}) = \mathbf{B}\boldsymbol{\theta}, \tag{2.8}$$

where $\mathbf{B} = \mathbf{B}(\boldsymbol{x})$ is a regression basis of dimension $n \times c$, which is constructed from the predictor $\boldsymbol{x}$, and $\boldsymbol{\theta}$ is the associated vector of regression coefficients of length c. Following the P-spline methodology, we can estimate $\boldsymbol{\theta}$ by minimizing the penalized sum of squares:

$$S_{\mathrm{P}} = (\boldsymbol{y} - \mathbf{B}\boldsymbol{\theta})' (\boldsymbol{y} - \mathbf{B}\boldsymbol{\theta}) + \boldsymbol{\theta}' \mathbf{P} \boldsymbol{\theta}, \tag{2.9}$$

where $\mathbf{P}$ is a penalty term that forces the coefficients to vary smoothly and, consequently, we obtain a smoothed curve. In order to control the amount of smoothness (that is, the trade-off between the model fit and the model smoothness), the term $\mathbf{P}$ will depend on a regularization parameter, called *smoothing parameter*, which we denote as λ. Then, for a given value of λ, the solution of the penalized least square problem given in Eq. (2.9) is:

$$\widehat{\boldsymbol{\theta}} = (\mathbf{B}'\mathbf{B} + \mathbf{P})^{-1} \mathbf{B}'\boldsymbol{y},$$

and the fitted values are given by:

$$\widehat{\boldsymbol{y}} = \mathbf{B} (\mathbf{B}'\mathbf{B} + \mathbf{P})^{-1} \mathbf{B}'\boldsymbol{y}. \tag{2.10}$$

The expression $(\mathbf{B}'\mathbf{B} + \mathbf{P})^{-1} \mathbf{B}'$ in Eq. (2.10) is called the *hat matrix* of the model,

which is denoted as $\mathbf{H} = (\mathbf{B}'\mathbf{B} + \mathbf{P})^{-1}\mathbf{B}'$. Notice that $\mathbf{H}$ is not idempotent (hence, it is not a projection matrix), but its form makes the smoothing method to be linear. Moreover, the hat matrix gives a measure of the *effective dimension* (ED) or the effective degrees of freedom of the model (Hastie and Tibshirani, 1990). This is calculated as the trace of $\mathbf{H}$, which can be efficiently computed as:

$$\text{ED} = \text{trace}(\mathbf{H}) = \text{trace}((\mathbf{B}'\mathbf{B} + \mathbf{P})^{-1}\mathbf{B}'\mathbf{B}).$$

Once we have briefly presented the P-spline methodology, we proceed to give details about the regression basis and penalties that we will use in the following chapters.

Regression basis

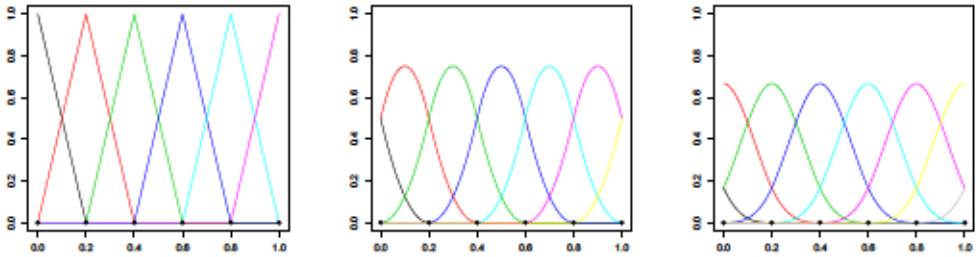
In the non-parametric literature, there exist several alternatives for the choice of the regression basis $\mathbf{B}$ in Eq. (2.8). For instance, we could use truncated p-th power functions (TPFs, Ruppert et al., 2003) that are characterized by their simplicity. Here we follow the proposal of Eilers and Marx (1996), in which a basis of B-splines is used (for more details about B-splines, see Dierckx, 1993; de Boor, 2001). B-splines are numerically stable and have nice properties and extensions than TPFs. Further comparisons between B-splines and TPFs are described in Eilers and Marx (2010).

In few words, a B-spline consists of polynomial pieces of the same degree, which are connected in a special way. Some general properties of a B-spline of degree p (extracted from Eilers and Marx, 1996) are:

- It consists of $p + 1$ polynomial pieces, each one of degree p, that join at p inner knots.
- At the joining points, derivatives up to order $p - 1$ are continuous.
- The B-spline is positive on a domain spanned by $p + 2$ knots, everywhere else it is zero.
- Except at the boundaries, it overlaps with exactly $2p$ polynomial pieces of its neighbours.

- At given x, $p+1$ B-splines are non-zero.

In practice it is usual to use polynomials of degree three (cubic B-splines), i.e, $p = 3$, and a moderate large number of equally-spaced knots (between 20 and 40). If we divide the domain $(x_{\min}, x_{\max})$ of $\boldsymbol{x}$ into k equal intervals by $k+1$ knots, each interval will be covered by $p+1$ B-splines of degree p. Therefore, the number of B-splines in the regression basis (i.e., the number of columns of matrix $\mathbf{B}$ in Eq. (2.8)) is $c = k + p$. Figure 2.1 shows examples of B-spline bases with $k = 5$ intervals and different degrees for the B-splines. In each panel, the B-splines have the same shape but shifted through the horizontal axis; a property that also holds at the boundaries.



(a) $p = 1$ (linear B-splines) (b) $p = 2$ (quadratic B-splines) (c) $p = 3$ (cubic B-splines)

Figure 2.1: B-spline bases of different orders of degree p, each one with $k = 5$ equally-spaced intervals.

Penalty matrix

Eilers and Marx (1996) propose to consider the penalty term $\mathbf{P}$ as a discrete matrix, which is based on finite differences of the regression coefficients associated to the B-spline basis. The penalty term $\mathbf{P}$ that they consider in Eq. (2.9) has the form:

$$\mathbf{P} = \lambda \mathbf{D}'\mathbf{D}, \tag{2.11}$$

where $\mathbf{D} = \mathbf{D}(q)$ is a q^{th} *order difference matrix* of dimension $(c-q) \times c$ and λ is the smoothing parameter.

The usual choice for the difference order q is two, although we can use higher or lower orders depending on the variability of the curve and the amount of noise on data. When $q = 2$, the matrix $\mathbf{D}$ has the following form:

$$\mathbf{D} = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & \cdots \\ 0 & 1 & -2 & 1 & 0 & \cdots \\ 0 & 0 & 1 & -2 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}_{(c-2)\times c},$$

and, for this case, the penalty term $\boldsymbol{\theta}'\mathbf{P}\boldsymbol{\theta}$ in Eq. (2.9) can be expressed as:

$$\boldsymbol{\theta}'\mathbf{P}\boldsymbol{\theta} = \lambda\boldsymbol{\theta}'\mathbf{D}'\mathbf{D}\boldsymbol{\theta} = \lambda\left((\theta_1 - 2\theta_2 + \theta_3)^2 + \cdots + (\theta_{c-2} - 2\theta_{c-1} + \theta_c)^2\right).$$

Under this framework, the ED associated to the P-spline model is given by:

$$\text{ED} = \text{trace}((\mathbf{B}'\mathbf{B} + \lambda\mathbf{D}'\mathbf{D})^{-1}\mathbf{B}'\mathbf{B}).$$

Here the value of the ED is determined by the size of the B-spline basis $\mathbf{B}$ used and the amount of smoothing given by λ in this sense: the values of ED varies between c (i.e., the number of columns of $\mathbf{B}$), when $\lambda \to 0$, and q (i.e., the order of the discrete penalty), when $\lambda \to \infty$.

Smoothing parameter selection

Previously we have assumed that the smoothing parameter λ is known. In order to select the optimal value for this parameter, Eilers and Marx (1996) suggest to minimize an information criterion of the form:

$$\text{IC} = \text{dev}(\boldsymbol{y}, \widehat{\boldsymbol{y}}) + \varpi \times \text{ED}, \tag{2.12}$$

where $\text{dev}(\boldsymbol{y}, \widehat{\boldsymbol{y}})$ denotes the *deviance* of the model, which measures the discrepancy between the fitted values of the model and data. For example, for the Gaussian case, $\text{dev}(\boldsymbol{y}, \widehat{\boldsymbol{y}}) = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$. For non-Gaussian data, the deviance is based on a generalization of this sum of squares and, depending on the distributional assumption, it will take a different form.

The term ϖ given in Eq. (2.12) can be viewed as a weight that penalizes

the effective dimension of the model. When $\varpi = 2$, the IC becomes the Akaike information criterion (AIC), and when $\varpi = \log(n)$, the IC corresponds to the Bayesian information criterion (BIC).

Alternatives to the use of information criteria for smoothing parameter selection are based on cross-validation methods. Ordinary cross-validation (CV) and generalized cross-validation (GCV) criteria are commonly used in this context (see Ruppert et al., 2003, for more details).

Further extensions

The P-spline methodology can be extended to the case of univariate non-Gaussian data under the GLM framework (Eilers and Marx, 1996). Moreover, they can be extended to the multidimensional case by means of tensor products of B-splines bases (Currie et al., 2006; Eilers et al., 2006). A nice feature of P-splines is that they are connected to mixed models (Currie and Durbán, 2002; Currie et al., 2006), leading to further insights, as well as to new methods for finding optimal values for the smoothing parameters in a multidimensional context. Within this mixed model framework, spatial and spatio-temporal models were developed and applied to health and environmental data (Lee and Durbán, 2009, 2011). More useful extensions and references about this flexible methodology can be found in Eilers et al. (2015).

2.1.3 The penalized composite link model

Once we have presented the Poisson CLM framework and the P-spline methodology, we can introduce the PCLM approach of Eilers (2007) for count data.

Suppose that we observe a vector of grouped counts $\boldsymbol{y}$, assumed Poisson distributed with mean vector $\boldsymbol{\mu}$. The aim now is to estimate the latent vector $\boldsymbol{\gamma}$ established in a Poisson CLM via P-spline methodology. In the model given in Eq. (2.3) we can redefine the linear predictor $\boldsymbol{\eta}$ as $\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\theta}$, where $\mathbf{B}$ is a B-spline basis of dimension $m \times c$ and $\boldsymbol{\theta}$ is its associated vector of regression coefficients of length c. Then the Poisson PCLM is given by:

$$\boldsymbol{\mu} = \mathbf{C}\boldsymbol{\gamma} = \mathbf{C}\exp(\mathbf{B}\boldsymbol{\theta}), \tag{2.13}$$

where the regression coefficients $\boldsymbol{\theta}$ are subject to the discrete penalization given in Eq. (2.11).

To estimate the model in Eq. (2.13), we can proceed in a similar fashion as in CLM estimation. Thus, let us consider the penalized log-likelihood:

$$\ell_{\mathrm{P}} = \ell - \frac{\lambda}{2}\boldsymbol{\theta}'\mathbf{D}'\mathbf{D}\boldsymbol{\theta}, \tag{2.14}$$

where ℓ is defined as in Eq. (2.4), but with $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)'$ defined now as in Eq. (2.13). Deriving Eq. (2.14) with respect to $\boldsymbol{\theta}$, and then equating the result to zero, we obtain the system of equations $\breve{\mathbf{B}}'(\boldsymbol{y} - \mathbf{C}\exp(\mathbf{B}\boldsymbol{\theta})) = \lambda\mathbf{D}'\mathbf{D}\boldsymbol{\theta}$, which leads to a penalized version of the IRWLS equations of a CLM (see Eq. (2.6)):

$$(\breve{\mathbf{B}}'\tilde{\mathbf{W}}\breve{\mathbf{B}} + \lambda\mathbf{D}'\mathbf{D})\widehat{\boldsymbol{\theta}} = \breve{\mathbf{B}}'\tilde{\mathbf{W}}(\tilde{\mathbf{W}}^{-1}(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}) + \breve{\mathbf{B}}\tilde{\boldsymbol{\theta}}), \tag{2.15}$$

where $\breve{\mathbf{B}} = \tilde{\mathbf{W}}^{-1}\mathbf{C}\tilde{\Gamma}\mathbf{B}$, $\tilde{\mathbf{W}} = \mathrm{diag}(\tilde{\boldsymbol{\mu}})$, $\tilde{\boldsymbol{\Gamma}} = \mathrm{diag}(\tilde{\gamma})$, with $\tilde{\boldsymbol{\mu}} = \mathbf{C}\tilde{\gamma} = \mathbf{C}\exp(\mathbf{B}\tilde{\boldsymbol{\theta}})$.

To find an optimal value for the smoothing parameter λ in this case, we can use an information criterion as in the P-spline methodology. Following the suggestion of Hastie and Tibshirani (1990), we take the trace of the following hat matrix $\mathbf{H}$ as the ED of the model in Eq. (2.13):

$$\mathbf{H} = \breve{\mathbf{B}}(\breve{\mathbf{B}}'\mathbf{W}\breve{\mathbf{B}} + \lambda\mathbf{D}'\mathbf{D})^{-1}\breve{\mathbf{B}}'\mathbf{W}, \tag{2.16}$$

which is implicit in Eq. (2.15). Thus, using the trace of hat matrix in Eq. (2.16) and the deviance of the model in the Poisson context:

$$\mathrm{dev}(\boldsymbol{y}, \widehat{\boldsymbol{y}}) = 2\sum_{i=1}^{n}\left(y_i \log\left(\frac{y_i}{\widehat{\mu}_i}\right) - (y_i - \widehat{\mu}_i)\right), \tag{2.17}$$

we can compute the required information criterion (see Eq. (2.12)). Then, a simple search algorithm for the smoothing parameter is sufficient: AIC (or BIC) is calculated for a fine grid of λ values (in a log scale) and its minimal value is determined over this grid. An R code for the previous estimation procedure can be found in (Rizzi et al., 2015, Appendix 2).

For illustration purposes, let us consider the number of deaths from respiratory diseases of American male population in January 1959, from ages 1 to 120 (see Currie et al., 2006, for more details about these data). Figure 2.2 shows the

counts per age-at-death (vertical lines) and the smooth trend that follow these counts ($g = 1$). If we artificially aggregate them into two, five, ten, and twenty age classes, and we apply the PCLM approach to these aggregated counts, we obtain the smooth colored curves of Figure 2.2 ($g = 2, 5, 10, 20$). The smooth curves for the cases $g = 2$, $g = 5$ and $g = 10$ are close to the smooth trend at the disaggregated scale, whereas the blue smooth curve for the case $g = 20$ departs from it (especially between 60 and 90 years old). This is because we have less precision when the aggregation level increases.

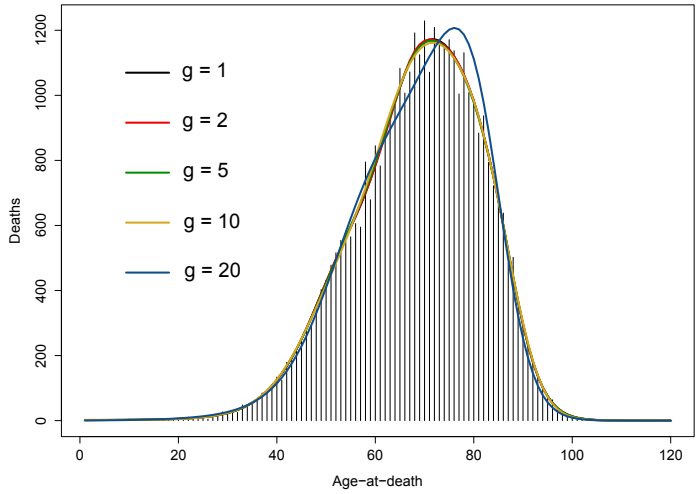


Figure 2.2: Death counts from respiratory disease of American population in January 1959, from ages 1 to 120 (vertical lines). The black curve represents the estimated trend based on the ungrouped data. The colored curves represent the estimated distributions using the PCLM approach, from different aggregations per g age classes, where g denotes the width of the groups.

The previous example shows the performance of the PCLM for different levels of aggregation. In Rizzi et al. (2016), the PCLM has been compared with other non-parametric methods for ungrouping aggregated count data. Specifically, the methods that are included in their comparison study are: a bootstrap kernel density estimator (Wang and Wertelecki, 2013), a piecewise cubic Hermite interpolating polynomial (Fritsch and Carlson, 1980), a spline interpolation with Hyman filter (Smith et al., 2004), and an iterated conditional expectation kernel density estimator using a local constant (Braun et al., 2005); all of them implemented in

R. Rizzi et al. (2016) concluded that these methods (including the PCLM) have a similar performance when the grouping scheme is relatively narrow, i.e., 5-year age classes; and with coarser age intervals, i.e., in the presence of open-ended age groups, the PCLM performs the best.

Once we have introduced the PCLM approach, in the next section we present our proposal: the penalized composite link mixed model. This new class of model allows the inclusion of specific random effects or further correlation structure if is necessary, and offers another alternative for the parameter estimation of the PCLM — avoiding the use of information criteria for smoothing parameter selection.

2.2 The composite link mixed model approach

In this section we present the composite link mixed model (CLMM). As we stated before, the CLMM can be seen as the reformulation of the PCLM given in Eq. (2.13) as a mixed model (in fact, as a generalized linear mixed model). To achieve this we follow the approach given in Currie and Durbán (2002) and Currie et al. (2006), where the B-spline basis $\mathbf{B}$ and the discrete penalty matrix $\mathbf{P}$ presented in Section 2.1 are used.

In order to introduce the mixed model formulation in our context, let first introduce a basic notion of the so-called linear mixed (effects) models.

2.2.1 Linear mixed models

A linear mixed model (LMM, Searle et al., 1992) is an extension of the linear regression model, which includes both fixed and random effects. Specifically, a basic LMM has the following structure:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \text{ with } \boldsymbol{\alpha} \backsim \mathcal{N}(\mathbf{0}, \mathbf{G}) \text{ and } \boldsymbol{\epsilon} \backsim \mathcal{N}(\mathbf{0}, \sigma^2_{\boldsymbol{\epsilon}}\mathbf{I}), \tag{2.18}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are the fixed and random effects coefficients, respectively, whereas $\mathbf{X}$ and $\mathbf{Z}$ are their corresponding design matrices. The random effects coefficients $\boldsymbol{\alpha}$ have covariance matrix $\mathbf{G}$ that depends on the variance $\sigma^2_{\boldsymbol{\alpha}}$ as $\mathbf{G} = \sigma^2_{\boldsymbol{\alpha}}\mathbf{R}$, where $\mathbf{R}$ is a positive semi-definite matrix. The error terms $\boldsymbol{\epsilon}$ are assumed independent and identically distributed with common variance $\sigma^2_{\boldsymbol{\epsilon}}$ (for a more general form or

further extensions of LMMs, see Searle et al., 1992; Pinheiro and Bates, 2000). The ML estimates for the fixed and random effects coefficients of the LMM in Eq. (2.18) are:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{y}, \\ \widehat{\boldsymbol{\alpha}} &= \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}), \end{aligned} \tag{2.19}$$

where $\mathbf{V} = \sigma_{\epsilon}^2\mathbf{I} + \mathbf{Z}\mathbf{G}\mathbf{Z}'$. Notice that matrix $\mathbf{V}$ includes both *variance components* σ_{ϵ}^2 and σ_{α}^2; the latter through $\mathbf{G}$.

The ML estimation, carried out to obtain the estimates given in Eq. (2.19), does not take into account the degrees of freedom used for estimating the fixed effects coefficients when estimating variance components, leading to biased estimates. To overcome this situation, we can use the so-called restricted (or residual) maximum likelihood estimation (Patterson and Thompson, 1971). Thus, the variance components can be estimated by maximizing the following restricted maximum log-likelihood (REML):

$$\ell_{\mathrm{R}}(\sigma_{\epsilon}^2, \sigma_{\alpha}^2) = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2}\boldsymbol{y}'(\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})\boldsymbol{y}.$$

There are several approaches to the formulation of P-splines as mixed models, which differ mainly in the regression basis used. For example, Brumback et al. (1999), Coull et al. (2001), and Wand (2003), have extended the model formulation into a mixed model, by using TPFs as regression bases. However, as it was previously mentioned, numerical properties of TPFs are poor compared to B-splines (Eilers et al., 2015). In the next subsection we describe the reparameterization that Currie and Durbán (2002) and Currie et al. (2006) propose to reformulate P-splines as mixed models, which will be subsequently used.

2.2.2 Mixed model formulation of P-splines

The reformulation proposed by Currie and Durbán (2002) and Currie et al. (2006) consists in to transform the B-spline basis $\mathbf{B}$ into a new model basis $[\mathbf{X} : \mathbf{Z}]$, and their associated vector of regression coefficients $\boldsymbol{\theta}$ into $(\boldsymbol{\beta}, \boldsymbol{\alpha})'$. This is achieved by considering an orthogonal transformation matrix $\mathbf{T}$, such that $\mathbf{B}\mathbf{T} = [\mathbf{X} : \mathbf{Z}]$ and $\mathbf{T}'\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})'$. The construction of matrix $\mathbf{T}$ is briefly described below.

Consider the singular value decomposition (SVD) of matrix $\mathbf{D}'\mathbf{D}$ in Eq. (2.11):

$$\mathbf{D}'\mathbf{D} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}',$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix that contains the eigenvalues of the SVD of $\mathbf{D}'\mathbf{D}$ (with q zero eigenvalues) and $\mathbf{U}$ is the corresponding matrix of eigenvectors. This latter matrix can decomposed as $\mathbf{U} = [\mathbf{U}_n : \mathbf{U}_s]$, where $\mathbf{U}_n$ and $\mathbf{U}_s$ are matrices of dimension $c \times q$ and $c \times (c-q)$ containing the eigenvectors associated to the null and non-null parts, respectively. Then, we can rewrite $\mathbf{D}'\mathbf{D}$ as follows:

$$\mathbf{D}'\mathbf{D} = [\mathbf{U}_n : \mathbf{U}_s] \begin{bmatrix} \mathbf{0}_q & \\ & \tilde{\boldsymbol{\Sigma}} \end{bmatrix} [\mathbf{U}_n : \mathbf{U}_s]', \tag{2.20}$$

where $\mathbf{0}_q$ denotes a square matrix of zeroes of dimension $q \times q$ and $\tilde{\boldsymbol{\Sigma}}$ is a diagonal matrix that contains the $(c-q)$ positive eigenvalues of the SVD of $\mathbf{D}'\mathbf{D}$. Thus, the required matrix $\mathbf{T}$ is defined as:

$$\mathbf{T} = [\mathbf{U}_n : \mathbf{U}_s]. \tag{2.21}$$

Given the transformation matrix in Eq. (2.21), it is easy to see that the fixed and random effects matrices $\mathbf{X}$ and $\mathbf{Z}$ can be obtained as:

$$\begin{aligned} \mathbf{X} &= \mathbf{B}\mathbf{U}_n, \\ \mathbf{Z} &= \mathbf{B}\mathbf{U}_s, \end{aligned} \tag{2.22}$$

and their associated coefficients can be written as $\boldsymbol{\beta} = \mathbf{U}_n'\boldsymbol{\theta}$ and $\boldsymbol{\alpha} = \mathbf{U}_s'\boldsymbol{\theta}$. This implies that the lengths of the vectors $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are q and $(c-q)$, respectively. Moreover, the new mixed model penalty matrix $\boldsymbol{\Upsilon}$ is obtained as:

$$\boldsymbol{\Upsilon} = \mathbf{T}'\mathbf{P}\mathbf{T} = \begin{bmatrix} \mathbf{0}_q & \\ & \mathbf{F} \end{bmatrix}, \text{ with } \mathbf{F} = \lambda\tilde{\boldsymbol{\Sigma}}, \tag{2.23}$$

and $\mathbf{P}$ defined as in Eq. (2.11). From Eq. (2.23), we can see that the fixed effects coefficients are unpenalized, whereas the random effects coefficients are penalized by the diagonal matrix $\mathbf{F}$. Then, the covariance matrix $\mathbf{G}$ associated to the random

effects can be written as:

$$\mathbf{G} = \sigma_\epsilon^2 \mathbf{F}^{-1}. \tag{2.24}$$

2.2.3 Composite link mixed models

Taking into account the reformulation given above, we can represent the PCLM in Eq. (2.13) as a mixed model.

As before, suppose that we observe a vector of grouped counts $\boldsymbol{y}$, assumed Poisson distributed with mean vector $\boldsymbol{\mu}$. In the model given in Eq. (2.13) we can redefine the linear predictor $\boldsymbol{\eta}$ as $\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha}$, where $\mathbf{X}$ and $\mathbf{Z}$ are the mixed model matrices defined in Eq. (2.22). Then, the Poisson CLMM is given by:

$$\boldsymbol{\mu} = \mathbf{C}\boldsymbol{\gamma} = \mathbf{C}(\boldsymbol{e}_\mathrm{f} * \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha})), \text{ with } \boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}), \tag{2.25}$$

where $\mathbf{G}$ is defined as in Eq. (2.24), with $\sigma_\epsilon^2 = 1$ (since we are in the Poisson case). Notice that we have included a vector $\boldsymbol{e}_\mathrm{f}$ of length m in Eq. (2.25). This vector allows to include exposures at the fine resolution where $\boldsymbol{\gamma}$ is defined, making possible to analyse rates instead of counts (if we only are interested in the analysis of counts, set $\boldsymbol{e}_\mathrm{f} = \mathbf{1}_m$).

If we take the composition matrix $\mathbf{C}$ in Eq. (2.25) as the identity matrix, then we have that $\boldsymbol{\mu} = \boldsymbol{\gamma}$ in Eq. (2.25). In such case, the CLMM approach is reduced to the P-spline methodology for a (Poisson) generalized linear mixed model (PGLMM) in a univariate setting.

Parameter estimation

Since the covariance matrix $\mathbf{G}$ in Eq. (2.25) is obtained from $\mathbf{F}$ in Eq. (2.23), it depends on the smoothing parameter λ that has to be estimated. As a consequence, the parameter estimation of the CLMM involves two interrelated stages: a) the estimation of the fixed and random effects coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ of the vector of latent expectations $\boldsymbol{\gamma}$; and b) the estimation of the smoothing parameter λ. For stage a), we use the penalized quasi-likelihood (PQL) approach (Breslow and Clayton, 1993), which is commonly used for the parameter estimation of GLMMs; and for stage b), we use the REML (Patterson and Thompson, 1971) as a numerical optimization criterion for smoothing parameter selection. Technical details of

these stages are derived below.

Consider the joint density function of $\boldsymbol{y}$ in the CLMM context:

$$f(\boldsymbol{y}|\boldsymbol{\alpha}) = \exp\left(\boldsymbol{y}' \log(\boldsymbol{\mu}) - \mathbf{1}_n'\boldsymbol{\mu} - \mathbf{1}_n' \log(\Gamma\left(\boldsymbol{y} + \mathbf{1}_n\right))\right), \tag{2.26}$$

where $\boldsymbol{\mu} = \mathbf{C}\boldsymbol{\gamma}$, $\boldsymbol{\gamma} = e_f * \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha})$, and $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}(\lambda))$. Taking into account Eq. (2.26) and for a given value of λ, we obtain estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ by maximizing the following penalized log-likelihood:

$$\ell_P = \log\left(f(\boldsymbol{y}|\boldsymbol{\alpha})\right) - \frac{1}{2}\boldsymbol{\alpha}'\mathbf{G}^{-1}\boldsymbol{\alpha}. \tag{2.27}$$

Differentiating Eq. (2.27) with respect to β_k and α_l, we obtain:

$$\frac{\partial \ell_P}{\partial \beta_k} = \sum_{i=1}^{n}\left((y_i - \mu_i)\frac{1}{\mu_i}\sum_{j=1}^{m} c_{ij}\gamma_j x_{jk}\right), \text{ for } k = 1, ..., q; \tag{2.28}$$

$$\frac{\partial \ell_P}{\partial \alpha_l} = \sum_{i=1}^{n}\left((y_i - \mu_i)\frac{1}{\mu_i}\sum_{j=1}^{m} c_{ij}\gamma_j z_{jl}\right) - \mathbf{G}_l^{-1}\boldsymbol{\alpha}, \text{ for } l = 1, ..., (c-q), \tag{2.29}$$

where $\mathbf{G}_l^{-1}$ denotes the l-th row of the matrix $\mathbf{G}^{-1}$. Writing $\frac{1}{\mu_i}\sum_{j=1}^{m} c_{ij}\gamma_j x_{jk}$ in Eq. (2.28) and $\frac{1}{\mu_i}\sum_{j=1}^{m} c_{ij}\gamma_j z_{jl}$ in Eq. (2.29) as $\breve{x}_{ik}$ and $\breve{z}_{il}$, respectively, and equating the expressions above to zero, we obtain:

$$\sum_{i=1}^{n}(y_i - \mu_i)\breve{x}_{ik} = 0, \text{ for } k = 1, ..., q;$$
$$\sum_{i=1}^{n}(y_i - \mu_i)\breve{z}_{il} = \mathbf{G}_l^{-1}\boldsymbol{\alpha}, \text{ for } l = 1, ..., (c-q).$$

Moreover, the equations above can be rewritten in matrix form as:

$$\breve{\mathbf{X}}'\left(\boldsymbol{y} - \boldsymbol{\mu}\right) = \mathbf{0}; \tag{2.30}$$

$$\breve{\mathbf{Z}}'\left(\boldsymbol{y} - \boldsymbol{\mu}\right) = \mathbf{G}^{-1}\boldsymbol{\alpha}, \tag{2.31}$$

where $\breve{\mathbf{X}} = \mathbf{W}^{-1}\mathbf{C}\Gamma\mathbf{X}$ and $\breve{\mathbf{Z}} = \mathbf{W}^{-1}\mathbf{C}\Gamma\mathbf{Z}$, with $\mathbf{W} = \text{diag}(\boldsymbol{\mu})$ and $\boldsymbol{\Gamma} = \text{diag}(\boldsymbol{\gamma})$. Defining the working vector:

$$\boldsymbol{z} = \breve{\mathbf{X}}\boldsymbol{\beta} + \breve{\mathbf{Z}}\boldsymbol{\alpha} + \mathbf{W}^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu}\right),$$

the solution of Eq. (2.30) and Eq.(2.31) via Fisher scoring algorithm (Green, 1987) can be expressed as the iterative solution of the system:

$$\begin{bmatrix} \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{X}} & \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}} \\ \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{X}} & \mathbf{G}^{-1}+\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} \breve{\mathbf{X}}'\mathbf{W}\boldsymbol{z} \\ \breve{\mathbf{Z}}'\mathbf{W}\boldsymbol{z} \end{bmatrix}. \tag{2.32}$$

Notice that the linear system given in Eq. (2.32) has exactly the same structure as that for a PGLMM (Lee, 2010). The difference is that in a PGLMM we would have $\mathbf{X}$ and $\mathbf{Z}$ while here $\breve{\mathbf{X}}$ and $\breve{\mathbf{Z}}$ appear. Thus $\breve{\mathbf{X}}$ and $\breve{\mathbf{Z}}$ are 'working' $\mathbf{X}$ and $\mathbf{Z}$ matrices, respectively. From Eq. (2.32) we obtain a modified version of the standard mixed model estimators (see Eq. (2.19)):

$$\widehat{\boldsymbol{\beta}} = (\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})^{-1}\breve{\mathbf{X}}'\mathbf{V}^{-1}\boldsymbol{z}, \tag{2.33}$$

$$\begin{aligned} \widehat{\boldsymbol{\alpha}} &= \mathbf{G}\breve{\mathbf{Z}}'\mathbf{V}^{-1}(\boldsymbol{z}-\breve{\mathbf{X}}\widehat{\boldsymbol{\beta}}) \\ &= \mathbf{G}\breve{\mathbf{Z}}'\mathbf{N}\boldsymbol{z}, \end{aligned} \tag{2.34}$$

where:

$$\mathbf{V} = \mathbf{W}^{-1}+\breve{\mathbf{Z}}\mathbf{G}\breve{\mathbf{Z}}', \tag{2.35}$$

$$\mathbf{N} = \mathbf{V}^{-1}-\mathbf{V}^{-1}\breve{\mathbf{X}}(\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})^{-1}\breve{\mathbf{X}}'\mathbf{V}^{-1}. \tag{2.36}$$

Conditioning on the estimates given in Eq. (2.33), the smoothing parameter λ can be estimated numerically by maximizing the approximate REML:

$$-\frac{1}{2}\log|\mathbf{V}|-\frac{1}{2}\log|\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}}|-\frac{1}{2}\boldsymbol{z}'(\mathbf{V}^{-1}-\mathbf{V}^{-1}\breve{\mathbf{X}}(\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})^{-1}\breve{\mathbf{X}}'\mathbf{V}^{-1})\boldsymbol{z}. \tag{2.37}$$

Therefore, the PQL solution is achieved by iteration between Eq. (2.33) and Eq. (2.37), until convergence. Notice that the terms $|\mathbf{V}|$ and $\mathbf{V}^{-1}$ appear in Eq. (2.37). From (2.35), we have they can be expressed as (see Searle et al., 1992, p. 453):

$$|\mathbf{V}| = |\mathbf{W}|^{-1}|\mathbf{G}|\,|\mathbf{G}^{-1}+\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}}|, \tag{2.38}$$

$$\mathbf{V}^{-1} = \mathbf{W}-\mathbf{W}\breve{\mathbf{Z}}(\mathbf{G}^{-1}+\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}})^{-1}\breve{\mathbf{Z}}'\mathbf{W}. \tag{2.39}$$

Approximate standard errors and effective dimension

Once the parameter values at convergence are obtained, we can derive standard errors for the mixed model estimators as shown in Lin and Zhang (1999), i.e., by approximating the covariance matrix of $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}})'$ by its Bayesian counterpart. This approximated covariance matrix is given by:

$$\mathbf{M} = \begin{bmatrix} \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{X}} & \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}} \\ \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{X}} & \mathbf{G}^{-1} + \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}} \end{bmatrix}^{-1}, \tag{2.40}$$

which corresponds to the inverse of the matrix on the left-hand side of Eq. (2.32). Thus we can obtain standard errors for $\widehat{\boldsymbol{\eta}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\boldsymbol{\alpha}}$ by taking the square root of the elements of $\mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{\eta}})$, which are obtained as:

$$\mathbb{V}\mathrm{ar}(\widehat{\eta}_j) = \mathrm{diag}([\mathbf{X} : \mathbf{Z}]\mathbf{M}[\mathbf{X} : \mathbf{Z}]')_{jj},$$

where $\mathbf{M}$ is defined in (2.40). Approximate standard errors for $\exp(\widehat{\boldsymbol{\eta}})$ can be derived by using the Delta method (see, e.g., Ver Hoef, 2012; Agresti, 2015):

$$\mathbb{V}\mathrm{ar}(\exp(\widehat{\eta}_j)) = \mathbb{V}\mathrm{ar}(\widehat{\eta}_j) \times (\exp(\widehat{\eta}_j))^2.$$

On the other hand, we can calculate the effective dimension of the CLMM (on the aggregated scale) as the trace of following hat matrix $\mathbf{H}$:

$$\mathbf{H} = [\breve{\mathbf{X}} : \breve{\mathbf{Z}}]\mathbf{M} \begin{bmatrix} \breve{\mathbf{X}}'\mathbf{W} \\ \breve{\mathbf{Z}}'\mathbf{W} \end{bmatrix},$$

with $\mathbf{M}$ defined in (2.40).

Illustration using mortality data

Here we illustrate the CLMM approach using a Canadian mortality dataset, which was obtained from the Human Mortality Database (HMD, http://www.mortality.org). The dataset contains the observed and expected number of deaths of Canadian female population by single year of age, from 1 to 105 years (i.e., $\boldsymbol{x} = (1, ..., 105)'$), for three selected years: 1960, 1990, and 2010. The corre-

sponding death rates (in log scale) are depicted in Figure 2.3.

To demonstrate the performance of the CLMM approach, we grouped the death counts, of each selected year, into 5-year age classes. To set up the CLMM formulation, we use 18 equally-spaced knots for the cubic B-spline basis and a second order penalty in each case. Also, we use the expected number of deaths as the vector of exposures e_f at the fine scale. The composition matrix for all the cases can be constructed in a similar fashion as in (2.2). Figure 2.3 shows the resulting CLMM estimates for the (log) death rates (solid lines), which are computed as $\mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\boldsymbol{\alpha}}$, and their associated 95% confidence intervals (dashed lines). The optimal values of the smoothing parameters for years 1960, 1990, and 2010 were 4.38, 12.64, and 3.88, respectively. In all the cases, the CLMM provides accurate results, except perhaps at younger and older ages. In fact, the confidence intervals are more wider in these parts.

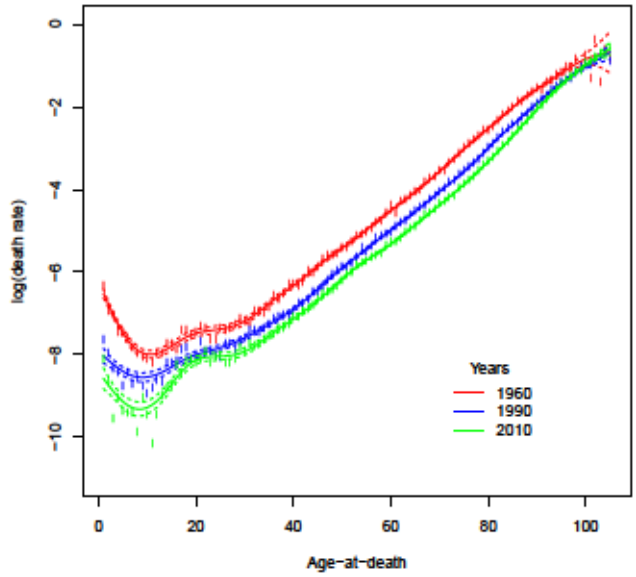


Figure 2.3: Raw female death rates (on log scale) in Canada (dot points), from ages 1 to 105 and for three selected years. The colored solid lines represent the estimated trends using the CLMM approach, from grouped counts in 5-year age classes. In each case, the dashed lines correspond to the approximate 95% confidence intervals of the estimated trend.

In the example provided above, the vector e_f is known in advance. If this were not the case, and only grouped exposures are available, we can use the CLMM approach to obtain exposure estimates at the required fine scale.

2.3 Multidimensional extension of CLMMs

In the previous section we have presented the CLMM approach for the univariate case, i.e., when count data are grouped into several classes along one dimension. But it may occur that these type of data are available in an aggregated form over multiple dimensions. For example, we can have a bivariate histogram of death counts, where the first dimension is formed by age classes and the second by calendar year intervals. In such case, it would be desirable to estimate the underlying mortality distribution by each single year old and each calendar year. Another example comes from disease mapping, where the observed and the expected numbers of deaths for particular disease are recorded at municipality level. A researcher could be interested in to analyse the spatial distribution of the mortality risk, but at a finer spatial resolution. The CLMM approach presented in Section 2.2 can be extended to handle such situations in an elegant way. In this new context, the vector γ will represent a smoothed surface at the desired fine scale.

In this section we will focus on the case when count data are available in multidimensional coarse grids, and will discuss the case of spatially aggregated count data in Chapter 3.

2.3.1 PCLMs for data with array structure

In order to present the extension of CLMMs to the multidimensional array case, let first introduce the extension of the PCLM approach in this context. For simplicity, we will illustrate the two-dimensional case below.

Let us consider $\boldsymbol{x}_1 = (x_{11}, ..., x_{1m_1})'$ and $\boldsymbol{x}_2 = (x_{21}, ..., x_{2m_2})'$ as two covariates defined at a fine scale. Suppose that we observe a vector of aggregated counts $\boldsymbol{y}$ of length $n = n_1 n_2$, with $n_d \leq m_d$, $d = 1, 2$, that have an array structure, i.e., $\boldsymbol{y} = \text{vec}(\mathbf{Y})$, where $\text{vec}(\cdot)$ denotes the vectorization operator and:

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n_2} \\ y_{21} & y_{22} & \cdots & y_{2n_2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n_1 1} & y_{n_1 2} & \cdots & y_{n_1 n_2} \end{bmatrix}.$$

Then, the PCLM given in Eq. (2.13) can be extended to the two-dimensional array

case by considering the matrix $\mathbf{B}$ as:

$$\mathbf{B} = \mathbf{B}_2 \otimes \mathbf{B}_1, \tag{2.41}$$

where each marginal B-spline basis $\mathbf{B}_d$ of dimension $m_d \times c_d$ is constructed from covariate $\boldsymbol{x}_d$, $d = 1, 2$; and the composition matrix $\mathbf{C}$ as:

$$\mathbf{C} = \mathbf{C}_2 \otimes \mathbf{C}_1, \tag{2.42}$$

where each marginal composition matrix $\mathbf{C}_d$ of dimension $n_d \times m_d$, $d = 1, 2$, reflects the aggregation process in each dimension. In Eq. (2.41) and Eq. (2.42) it appears the matrix operator $\otimes$, which denotes the Kronecker product of two matrices. Regarding the penalty matrix $\mathbf{P}$, it can be generalized to the two-dimensional case as:

$$\mathbf{P} = \lambda_1 \mathbf{I}_{c_2} \otimes \mathbf{D}_1' \mathbf{D}_1 + \lambda_2 \mathbf{D}_2' \mathbf{D}_2 \otimes \mathbf{I}_{c_1}, \tag{2.43}$$

where λ_d is a smoothing parameter that controls the amount of the smoothness along the covariate $\boldsymbol{x}_d$, and $\mathbf{D}_d$ is the q_d^{th} order difference matrix, $d = 1, 2$. The penalty matrix in Eq. (2.43) is anisotropic, since it considers a different amount of smoothing in each dimension. Moreover, this matrix can be written in a compact notation as:

$$\mathbf{P} = \lambda_2 \mathbf{D}_2' \mathbf{D}_2 \oplus \lambda_1 \mathbf{D}_1' \mathbf{D}_1,$$

where the matrix operator $\oplus$ denotes the Kronecker sum of two matrices.

Notice that, in this new context, the vector of regression coefficients $\boldsymbol{\theta}$ of length $c_1 c_2$ can be arranged into a matrix $\boldsymbol{\Theta}$ of dimension $c_1 \times c_2$. Thus, we have that $\boldsymbol{\theta} = \text{vec}(\boldsymbol{\Theta})$, where:

$$\boldsymbol{\Theta} = \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1c_2} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2c_2} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{c_1 1} & \theta_{c_1 2} & \cdots & \theta_{c_1 c_2} \end{bmatrix}.$$

Therefore, the expression $\mathbf{B}\boldsymbol{\theta}$ (with $\mathbf{B}$ defined as in Eq. (2.41)) of the model given in Eq. (2.3) can be computed as:

$$(\mathbf{B}_2 \otimes \mathbf{B}_1)\boldsymbol{\theta} = \text{vec}(\mathbf{B}_1 \boldsymbol{\Theta} \mathbf{B}_2'). \tag{2.44}$$

The expression in right-hand side of Eq. (2.44) avoids the computation of matrix $\mathbf{B}$ in Eq. (2.41). Also, the number of multiplications on the right-hand side is much less than that on the left hand side.

The estimation of the two-dimensional PCLM is done using the same methodology that we showed in Section 2.1, where now the strategy is to evaluate the information criterion for a fine grid of values of the smoothing parameters. However, since Kronecker products of matrices are involved here, the estimation process is susceptible to run into problems in terms of storage and computational burden. A solution for this case is the appropriate use of an arithmetic of arrays provided by Currie et al. (2006) and Eilers et al. (2006). These algorithms are referred as generalized linear array models, or GLAMs, since they are developed under the GLM framework. The GLAM methods were inspired to provide efficient array computations as in Eq. (2.44). Since we are working in a mixed model framework, we will show the use of GLAM algorithms in the CLMM context later.

Three-dimensional case

The extension of the PCLM approach to more than two dimensions is straightforward. For example, for the three-dimension case, the grouped counts $\boldsymbol{y}$ can be arranged in a three-dimensional array of dimension $n_1 \times n_2 \times n_3$. The corresponding regression basis $\mathbf{B}$ for the PCLM is:

$$\mathbf{B} = \mathbf{B}_3 \otimes \mathbf{B}_2 \otimes \mathbf{B}_1, \tag{2.45}$$

where each $\mathbf{B}_d$ is a marginal B-spline basis of dimension $m_d \times c_d$ constructed from a covariate $\boldsymbol{x}_d$ defined at fine scale, for $d = 1, 2, 3$. Thus, the dimension of the matrix in Eq. (2.45) is $m_1 m_2 m_3 \times c_1 c_2 c_3$, where m_d is the length of the vector $\boldsymbol{x}_d$. The composition matrix for this case is:

$$\mathbf{C} = \mathbf{C}_3 \otimes \mathbf{C}_2 \otimes \mathbf{C}_1, \tag{2.46}$$

and the three-dimensional penalty matrix that penalizes the regression coefficients $\boldsymbol{\theta}$ is:

$$\mathbf{P} = \lambda_1 \mathbf{D}_1' \mathbf{D}_1 \otimes \mathbf{I}_{c_2} \otimes \mathbf{I}_{c_3} + \lambda_2 \mathbf{I}_{c_1} \otimes \mathbf{D}_2' \mathbf{D}_2 \otimes \mathbf{I}_{c_3} + \lambda_3 \mathbf{I}_{c_1} \otimes \mathbf{I}_{c_2} \otimes \mathbf{D}_3' \mathbf{D}_3, \tag{2.47}$$

or, expressed in terms of Kronecker sums:

$$\mathbf{P} = \lambda_3 \mathbf{D}_3' \mathbf{D}_3 \oplus \lambda_2 \mathbf{D}_2' \mathbf{D}_2 \oplus \lambda_1 \mathbf{D}_1' \mathbf{D}_1.$$

2.3.2 Multidimensional mixed model formulation of P-splines

In order to extend the CLMM to the multidimensional array case, we use the multidimensional mixed formulation of P-splines. The formulation for the two-dimensional case is briefly described below. For more details, see Lee (2010).

The aim is to find a transformation matrix $\mathbf{T}$ such that we can reparameterize the regression basis $\mathbf{B}$ in Eq. (2.41) and its associated regression coefficients $\boldsymbol{\theta}$ as:

$$\mathbf{B} \to [\mathbf{X} : \mathbf{Z}] \quad \text{and} \quad \boldsymbol{\theta} \to (\boldsymbol{\beta}, \boldsymbol{\alpha})'.$$

For that, we consider the SVD of the marginal penalty $\mathbf{P}_d = \mathbf{D}_d' \mathbf{D}_d$ that are involved in Eq. (2.43):

$$\mathbf{P}_d = \mathbf{U}_d \boldsymbol{\Sigma}_d \mathbf{U}_d',$$

where $\boldsymbol{\Sigma}_d$ is a diagonal matrix that contains the eigenvalues of the SVD of $\mathbf{P}_d$ and $\mathbf{U}_d$ is the corresponding matrix of eigenvectors, for $d = 1, 2$. As in the univariate case, each matrix $\mathbf{P}_d$ can be decomposed as in Eq. (2.20), where now $\mathbf{U}_{dn}$ and $\mathbf{U}_{ds}$ are matrices containing the eigenvectors associated to the null and non-null parts, respectively, and $\tilde{\boldsymbol{\Sigma}}_d$ has $(c_d - q_d)$ positive eigenvalues, for $d = 1, 2$. Therefore, we can define a suitable transformation matrix $\mathbf{T}$ as:

$$\mathbf{T} = [\underbrace{\mathbf{U}_{2n} \otimes \mathbf{U}_{1n}}_{\mathbf{T}_n} : \underbrace{\mathbf{U}_{2s} \otimes \mathbf{U}_{1n} : \mathbf{U}_{2n} \otimes \mathbf{U}_{1s} : \mathbf{U}_{2s} \otimes \mathbf{U}_{1s}}_{\mathbf{T}_s}], \tag{2.48}$$

which is obtained by reordering the block matrices of the matrix $[\mathbf{U}_{2n} : \mathbf{U}_{2s}] \otimes [\mathbf{U}_{1n} : \mathbf{U}_{1s}]$. Then, given the transformation matrix in Eq. (2.48), the mixed model matrices for the two-dimensional case are obtained as:

$$\mathbf{X} = \mathbf{B}\mathbf{T}_n = (\mathbf{B}_2 \otimes \mathbf{B}_1)(\mathbf{U}_{2n} \otimes \mathbf{U}_{1n}) = \mathbf{B}_2\mathbf{U}_{2n} \otimes \mathbf{B}_1\mathbf{U}_{1n},$$

and

$$\begin{aligned}\mathbf{Z} &= \mathbf{B}\mathbf{T}_s \\ &= (\mathbf{B}_2 \otimes \mathbf{B}_1)\left[\mathbf{U}_{2s} \otimes \mathbf{U}_{1n} : \mathbf{U}_{2n} \otimes \mathbf{U}_{1s} : \mathbf{U}_{2s} \otimes \mathbf{U}_{1s}\right] \\ &= \left[\mathbf{B}_2\mathbf{U}_{2s} \otimes \mathbf{B}_1\mathbf{U}_{1n} : \mathbf{B}_2\mathbf{U}_{2n} \otimes \mathbf{B}_1\mathbf{U}_{1s} : \mathbf{B}_2\mathbf{U}_{2s} \otimes \mathbf{B}_1\mathbf{U}_{1s}\right].\end{aligned}$$

Denoting the matrices $\mathbf{X}_d = \mathbf{B}_d\mathbf{U}_{dn}$ and $\mathbf{Z}_d = \mathbf{B}_d\mathbf{U}_{ds}$ $(d = 1, 2)$, the previous mixed model matrices can be expressed as:

$$\begin{aligned}\mathbf{X} &= \mathbf{X}_2 \otimes \mathbf{X}_1, \\ \mathbf{Z} &= [\mathbf{Z}_2 \otimes \mathbf{X}_1 : \mathbf{X}_2 \otimes \mathbf{Z}_1 : \mathbf{Z}_2 \otimes \mathbf{Z}_1].\end{aligned} \tag{2.49}$$

The new mixed model coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are obtained from $\boldsymbol{\theta}$ as $\boldsymbol{\beta} = \mathbf{T}_n'\boldsymbol{\theta}$ and $\boldsymbol{\alpha} = \mathbf{T}_s'\boldsymbol{\theta}$.

Given the transformation matrix $\mathbf{T}$ in Eq. (2.48) and the two-dimensional penalty matrix defined in Eq. (2.43), the new two-dimensional mixed model penalty matrix $\boldsymbol{\Upsilon}$ is given as in Eq. (2.23), but with $q = q_1q_2$ and $\mathbf{F}$ defined as the block-diagonal matrix:

$$\mathbf{F} = \begin{bmatrix} \lambda_2\tilde{\boldsymbol{\Sigma}}_2 \otimes \mathbf{I}_{q_1} & & \\ & \lambda_1\mathbf{I}_{q_2} \otimes \tilde{\boldsymbol{\Sigma}}_1 & \\ & & \lambda_2\tilde{\boldsymbol{\Sigma}}_2 \otimes \mathbf{I}_{c_1-q_1} + \lambda_1\mathbf{I}_{c_2-q_2} \otimes \tilde{\boldsymbol{\Sigma}}_1 \end{bmatrix}, \tag{2.50}$$

where the matrices $\tilde{\boldsymbol{\Sigma}}_d$ $(d = 1, 2)$ were defined above.

Using the previous mixed model representation, we can extend the CLMM given in Eq. (2.25) to the two-dimensional array case, by considering now the mixed model matrices $\mathbf{X}$ and $\mathbf{Z}$ defined in Eq. (2.49), the composition matrix given in Eq. (2.42), and the covariance matrix $\mathbf{G}$ as in Eq. (2.24) but with $\mathbf{F}$ defined as in (2.50). The estimation procedure can be carried out as it was shown in Section 2.2. In the next section we will detail array computations for the two-dimensional CLMM using GLAM algorithms.

Three-dimensional case

We can also extend the mixed model formulation of P-splines to the multidimensional case. For example, for the three-dimensional case, the transformation matrix $\mathbf{T}$ can be defined as:

$$\mathbf{T} = [\mathbf{U}_{3n} : \mathbf{U}_{3s}] \otimes [\mathbf{U}_{2n} : \mathbf{U}_{2s}] \otimes [\mathbf{U}_{1n} : \mathbf{U}_{1s}]. \tag{2.51}$$

As in the two-dimensional case, we reorder the block matrices in Eq. (2.51) into two sub-blocks as $\mathbf{T} = [\mathbf{T}_n : \mathbf{T}_s]$, where:

$$\begin{aligned}
\mathbf{T}_n &= \mathbf{U}_{3n} \otimes \mathbf{U}_{2n} \otimes \mathbf{U}_{1n}, \\
\mathbf{T}_s &= [\, \mathbf{U}_{3s} \otimes \mathbf{U}_{2n} \otimes \mathbf{U}_{1n} : \mathbf{U}_{3n} \otimes \mathbf{U}_{2s} \otimes \mathbf{U}_{1n} : \mathbf{U}_{3n} \otimes \mathbf{U}_{2n} \otimes \mathbf{U}_{1s} : \\
&\qquad \mathbf{U}_{3s} \otimes \mathbf{U}_{2s} \otimes \mathbf{U}_{1n} : \mathbf{U}_{3s} \otimes \mathbf{U}_{2n} \otimes \mathbf{U}_{1s} : \mathbf{U}_{3n} \otimes \mathbf{U}_{2s} \otimes \mathbf{U}_{1s} : \mathbf{U}_{3s} \otimes \mathbf{U}_{2s} \otimes \mathbf{U}_{1s}].
\end{aligned}$$

Then, the three-dimensional mixed model matrices for the CLMM are given by:

$$\begin{aligned}
\mathbf{X} &= \mathbf{X}_3 \otimes \mathbf{X}_2 \otimes \mathbf{X}_1, \\
\mathbf{Z} &= [\, \mathbf{Z}_3 \otimes \mathbf{X}_2 \otimes \mathbf{X}_1 : \mathbf{X}_3 \otimes \mathbf{Z}_2 \otimes \mathbf{X}_1 : \mathbf{X}_3 \otimes \mathbf{X}_2 \otimes \mathbf{Z}_1 : \\
&\qquad \mathbf{Z}_3 \otimes \mathbf{Z}_2 \otimes \mathbf{X}_1 : \mathbf{Z}_3 \otimes \mathbf{X}_2 \otimes \mathbf{Z}_1 : \mathbf{X}_3 \otimes \mathbf{Z}_2 \otimes \mathbf{Z}_1 : \mathbf{Z}_3 \otimes \mathbf{Z}_2 \otimes \mathbf{Z}_1],
\end{aligned} \tag{2.52}$$

where $\mathbf{X}_d = \mathbf{B}_d\mathbf{U}_{dn}$ and $\mathbf{Z}_d = \mathbf{B}_d\mathbf{U}_{ds}$, for $d = 1, 2, 3$.

The block diagonal matrix $\mathbf{F}$ for the three-dimensional case is given by:

$$\begin{aligned}
\mathbf{F} = \text{blockdiag}\, (&\lambda_3\mathbf{F}_{3u}, \lambda_2\mathbf{F}_{2u}, \lambda_1\mathbf{F}_{1u}, \\
&\lambda_2\mathbf{F}_{22} + \lambda_3\mathbf{F}_{32}, \lambda_1\mathbf{F}_{12} + \lambda_3\mathbf{F}_{31}, \lambda_1\mathbf{F}_{11} + \lambda_2\mathbf{F}_{21}, \\
&\lambda_1\mathbf{F}_{1t} + \lambda_2\mathbf{F}_{2t} + \lambda_3\mathbf{F}_{3t})\,,
\end{aligned} \tag{2.53}$$

where:

$$\begin{aligned}
&\mathbf{F}_{1u} = \mathbf{I}_{q_3} \otimes \mathbf{I}_{q_2} \otimes \tilde{\mathbf{\Sigma}}_1, && \mathbf{F}_{2u} = \mathbf{I}_{q_3} \otimes \tilde{\mathbf{\Sigma}}_2 \otimes \mathbf{I}_{q_1}, && \mathbf{F}_{3u} = \tilde{\mathbf{\Sigma}}_3 \otimes \mathbf{I}_{q_2} \otimes \mathbf{I}_{q_1}, \\
&\mathbf{F}_{11} = \mathbf{I}_{q_3} \otimes \mathbf{I}_{c_2-q_2} \otimes \tilde{\mathbf{\Sigma}}_1, && \mathbf{F}_{12} = \mathbf{I}_{c_3-q_3} \otimes \mathbf{I}_{q_2} \otimes \tilde{\mathbf{\Sigma}}_1, && \mathbf{F}_{21} = \mathbf{I}_{q_3} \otimes \tilde{\mathbf{\Sigma}}_2 \otimes \mathbf{I}_{c_1-q_1}, \\
&\mathbf{F}_{22} = \mathbf{I}_{c_3-q_3} \otimes \tilde{\mathbf{\Sigma}}_2 \otimes \mathbf{I}_{q_1}, && \mathbf{F}_{31} = \tilde{\mathbf{\Sigma}}_3 \otimes \mathbf{I}_{q_2} \otimes \mathbf{I}_{c_1-q_1}, && \mathbf{F}_{32} = \tilde{\mathbf{\Sigma}}_3 \otimes \mathbf{I}_{c_2-q_2} \otimes \mathbf{I}_{q_1}, \\
&\mathbf{F}_{1t} = \mathbf{I}_{c_3-q_3} \otimes \mathbf{I}_{c_2-q_2} \otimes \tilde{\mathbf{\Sigma}}_1, && \mathbf{F}_{2t} = \mathbf{I}_{c_3-q_3} \otimes \tilde{\mathbf{\Sigma}}_2 \otimes \mathbf{I}_{c_1-q_1}, && \mathbf{F}_{3t} = \tilde{\mathbf{\Sigma}}_3 \otimes \mathbf{I}_{c_2-q_2} \otimes \mathbf{I}_{c_1-q_1}.
\end{aligned}$$

The matrix $\mathbf{F}$ in (2.53) will be used in Chapter 4, where we will analyse count data that are spatio-temporally aggregated.

2.3.3 Array methods for multidimensional CLMMs

When we are dealing with the estimation of the underlying distribution in several dimensions, we are susceptible to encounter problems with storage and computational time. In the case of data arranged in multidimensional grids, it is possible to circumvent these problems using the GLAM algorithms developed by Currie et al. (2006) and Eilers et al. (2006). In this section we show the use of GLAM algorithms in the PCLMM context, when the aggregated data have array structure. In Section 2.2, we proposed the use of the restricted maximum log-likelihood (REML) for the estimation of the smoothing parameters. Given (2.37) and the definitions of $\mathbf{V}$, $|\mathbf{V}|$ and $\mathbf{V}^{-1}$ in Eq. (2.35), Eq. (2.38), and Eq. (2.39), we can use the GLAM algorithms for a fast and efficient computation of the matrix cross-products: $\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}}$, $\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}}$, $\breve{\mathbf{X}}'\mathbf{W}\boldsymbol{z}$, $\breve{\mathbf{Z}}'\mathbf{W}\boldsymbol{z}$, etc., and estimate the smoothing parameters by REML.

To illustrate the implementation of the GLAM algorithms, we divide the expression for REML in four parts as:

$$-\frac{1}{2}\underbrace{\log|\mathbf{V}|}_{\text{part I}}-\frac{1}{2}\underbrace{\log|\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}}|}_{\text{part II}}-\frac{1}{2}(\underbrace{\boldsymbol{z}'\mathbf{V}^{-1}\boldsymbol{z}}_{\text{part III}}-\underbrace{\boldsymbol{z}'\mathbf{V}^{-1}\breve{\mathbf{X}}(\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})^{-1}\breve{\mathbf{X}}'\mathbf{V}^{-1}\boldsymbol{z}}_{\text{part IV}}).$$

Here we use some GLAM notation and definitions proposed by Currie et al. (2006) and Eilers et al. (2006), as for example, the row tensor of two matrices, $\mathcal{G}$, and the rotated $\mathcal{H}$-transform of an array by a matrix, ρ (for their definitions, see Appendix A).

Part I: Array computation of $\log|\mathbf{V}|$

Given the covariance matrix $\mathbf{G}=\sigma_\epsilon^2\mathbf{F}^{-1}$, with $\sigma_\epsilon^2=1$ (Poisson case) and $\mathbf{F}$ defined in (2.50), and considering Eq. (2.38), the term $\log|\mathbf{V}|$ can be written as:

$$\log|\mathbf{V}| = -\log|\mathbf{W}| + \log|\mathbf{F}^{-1}| + \log|\mathbf{F}+\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}}|. \tag{2.54}$$

Since $\mathbf{W}$ is a diagonal matrix and $\mathbf{F}$ is a block-diagonal matrix, the first two terms in Eq. (2.54) are calculated as $-\log|\mathbf{W}| = -\sum\log(\mu_i)$ and $\log|\mathbf{F}^{-1}| =$

$-\sum \log(\mathbf{F}_{ii})$, where $\mathbf{F}_{ii}$ denote the diagonal elements of $\mathbf{F}$.

For the computation of $\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}}$ in Eq. (2.54), notice that we can reduce this expression as:

$$\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}} = (\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z})'\mathbf{W}^{-1}(\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z}). \tag{2.55}$$

Since the composition matrix $\mathbf{C}$ is given by $\mathbf{C} = \mathbf{C}_2 \otimes \mathbf{C}_1$ and the model matrix $\mathbf{Z}$ can be rewritten as $\mathbf{Z} = [\mathbf{Z}_2 \otimes \mathbf{X}_1 : \tilde{\mathbf{Z}}_2 \otimes \mathbf{Z}_1]$, where $\tilde{\mathbf{Z}}_2 = \mathbf{X}_2 \otimes \mathbf{Z}_2$, the product of matrices $\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z}$ in Eq. (2.55) can be computed as:

$$\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z} \equiv [\rho(\mathcal{G}(\mathbf{Z}_2, \mathbf{C}_2')', \rho(\mathcal{G}(\mathbf{X}_1, \mathbf{C}_1')', \tilde{\boldsymbol{\Gamma}})) : \rho(\mathcal{G}(\tilde{\mathbf{Z}}_2, \mathbf{C}_2')', \rho(\mathcal{G}(\mathbf{Z}_1, \mathbf{C}_1')', \tilde{\boldsymbol{\Gamma}}))], \tag{2.56}$$

where $\tilde{\boldsymbol{\Gamma}}$ is a matrix of dimension $m_1 \times m_2$, whose entries are the elements of the diagonal of $\boldsymbol{\Gamma}$, that is, $\text{vec}(\tilde{\boldsymbol{\Gamma}}) = \boldsymbol{\gamma}$. The symbol $\equiv$ means that both sides have the same elements but in a different order.

Part II: Array computation of $\log|\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}}|$

Using Eq. (2.39), we can rewrite $\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}}$ as:

$$\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}} = \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{X}} - \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}}(\mathbf{F} + \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}})^{-1}\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{X}}. \tag{2.57}$$

Since $\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}}$ was previously calculated and $\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{X}} = (\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}})'$, we only need to compute the expressions $\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{X}}$ and $\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}}$ in Eq. (2.57). Notice that we can reduce them as:

$$\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{X}} = (\mathbf{C}\boldsymbol{\Gamma}\mathbf{X})'\mathbf{W}^{-1}(\mathbf{C}\boldsymbol{\Gamma}\mathbf{X}), \tag{2.58}$$

$$\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}} = (\mathbf{C}\boldsymbol{\Gamma}\mathbf{X})'\mathbf{W}^{-1}(\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z}). \tag{2.59}$$

where the expression $\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z}$ was calculated in Eq. (2.56). Considering the mixed model matrix $\mathbf{X} = \mathbf{X}_2 \otimes \mathbf{X}_1$, the expression $\mathbf{C}\boldsymbol{\Gamma}\mathbf{X}$, which appears in Eq. (2.58) and Eq. (2.59), can be computed as:

$$\mathbf{C}\boldsymbol{\Gamma}\mathbf{X} \equiv \rho(\mathcal{G}(\mathbf{X}_2, \mathbf{C}_2')', \rho(\mathcal{G}(\mathbf{X}_1, \mathbf{C}_1')', \tilde{\boldsymbol{\Gamma}})), \tag{2.60}$$

with $\tilde{\boldsymbol{\Gamma}}$ defined above.

Part III: Array computation of $z'\mathbf{V}^{-1}z$

Given Eq. (2.39), we can write $z'\mathbf{V}^{-1}z$ as:

$$z'\mathbf{V}^{-1}z = z'\mathbf{W}z - z'\mathbf{W}\breve{\mathbf{Z}}(\mathbf{F} + \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}})^{-1}\breve{\mathbf{Z}}'\mathbf{W}z, \tag{2.61}$$

where $z'\mathbf{W}z$ is calculated as $\sum \mu_i z_i^2$. We can rewrite the expression $z'\mathbf{W}\breve{\mathbf{Z}}$ in Eq. (2.61) as:

$$z'\mathbf{W}\breve{\mathbf{Z}} = z'\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z},$$

where $\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z}$ was calculated in Eq. (2.56).

Part IV: Array computation of $z'\mathbf{V}^{-1}\breve{\mathbf{X}}(\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})^{-1}\breve{\mathbf{X}}'\mathbf{V}^{-1}z$

We have shown how to compute $\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}}$ in Eq. (2.57). Thus, we only need to compute $z'\mathbf{V}^{-1}\breve{\mathbf{X}}$ (since $\breve{\mathbf{X}}'\mathbf{V}^{-1}z = (z'\mathbf{V}^{-1}\breve{\mathbf{X}})'$). Given Eq. (2.39), we can write $z'\mathbf{V}^{-1}\breve{\mathbf{X}}$ as:

$$z'\mathbf{V}^{-1}\breve{\mathbf{X}} = z'\mathbf{W}\breve{\mathbf{X}} - z'\mathbf{W}\breve{\mathbf{Z}}(\mathbf{F} + \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}})^{-1}\breve{\mathbf{Z}}'\mathbf{W}\mathbf{X}, \tag{2.62}$$

where all the quantities were computed previously, except $z'\mathbf{W}\breve{\mathbf{X}}$, which is computed as:

$$z'\mathbf{W}\breve{\mathbf{X}} = z'\mathbf{C}\boldsymbol{\Gamma}\mathbf{X},$$

where $\mathbf{C}\boldsymbol{\Gamma}\mathbf{X}$ was calculated in Eq. (2.60).

2.3.4 Illustrations

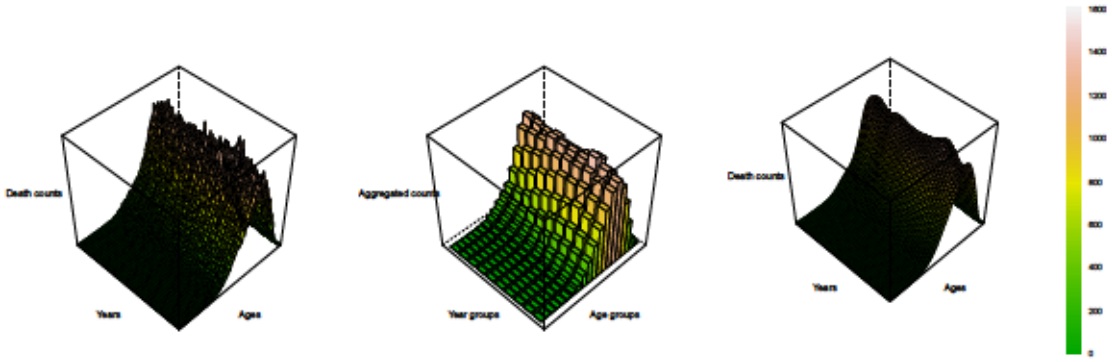
Here we illustrate the CLMM approach for coarsely grouped data with array structure. For the two-dimensional case, we use a dataset related with American male deaths by respiratory diseases (indexed by age and year at death); and for the three-dimensional case, we use a Canadian fertility dataset, which is recorded by age of the mother, calendar years, and birth order.

Deaths by respiratory diseases

Consider the death counts by respiratory diseases of American males from ages 1 to 100, and from 1959 to 1998 (for more details about this data, see Currie et al.,

2006). These raw data are displayed in Figure 2.4a. Suppose that we observe aggregated death counts, recorded in five-year age and four-year classes, instead of the previous raw data. Figure 2.4b shows the bivariate histogram for these aggregated counts, which is formed by 200 classes (that is, the resulting product of the 20 age groups and 10 year groups).

In order to estimate the underlying distribution behind these aggregated data, we apply the CLMM approach (for array data) described previously. In this case, $\boldsymbol{x}_1 = (1, ..., 100)'$ and $\boldsymbol{x}_2 = (1959, ..., 1998)'$ are the vectors of ages and years at the fine resolution, and $\mathbf{C} = \mathbf{C}_2 \otimes \mathbf{C}_1$, where $\mathbf{C}_1$ and $\mathbf{C}_2$ are the (marginal) composition matrices for ages and years, of dimensions 20×100 and 10×40, respectively. Figure 2.4c shows the smoothed bivariate distribution obtained from the CLMM approach, where number of equally-spaced knots used for $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are 25 and 10, respectively. We observe that the smoothed distribution closely follows the bivariate trend displayed by the original raw data. This is due in part by the levels of aggregation of each dimension. In general, the smoothed CLMM distribution will lose precision, if we observe wide classes at the edge of the histogram. Since we are only considering death counts, a way to improve the description of the mortality is considering a vector of exposures.



(a) Raw counts (b) Grouped counts (c) Estimated CLMM distribution

Figure 2.4: American male deaths by respiratory diseases during the period 1959-1998, from ages 1 to 100. The left and middle panels represent these deaths as totals of one-year age/one-year classes and five-year age/four-year classes respectively. The right panel shows the estimated distribution using Poisson CLMM approach for array data.

Considering the array methods described previously into the iterative procedure (to obtain the estimated distribution depicted in Figure 2.4c), the resulting computing time took about 61.840 seconds (Intel® Core™ i7, 1.80 GHz, Windows 8.1). On the other hand, if we disregard the use of these array methods, the computing time took about 221.360 seconds; that is, for this case, the computing time was reduced in about 3.6 times. This shows the usefulness of the adapted GLAM algorithms developed in this section, for CLMM estimation, in terms of computational speed.

Fertility rates collected by age, year, and birth order

To illustrate the CLMM for the three-dimensional, we use a Canadian fertility dataset (downloaded from the Human Fertility database, HFD, http://www.humanfertility.org/) that consists of birth counts and age-specific fertility rates recorded by age (which varies from 15 to 54 years), calendar year (from 1944 to 2009), and birth order (from 1st to 4th). The female population exposure can be then derived using these data. To show how our methodology works, we have grouped the birth counts and the exposures into 5-year age classes, and calendar year classes of different lengths (starting with four classes of length 10, followed by five classes of length 5). The left panels of Figure 2.5a, Figure 2.5b, Figure 2.5c, and Figure 2.5d show the fertility rates resulting from this aggregation for the 1st, 2nd, 3rd, and 4th birth order, respectively.

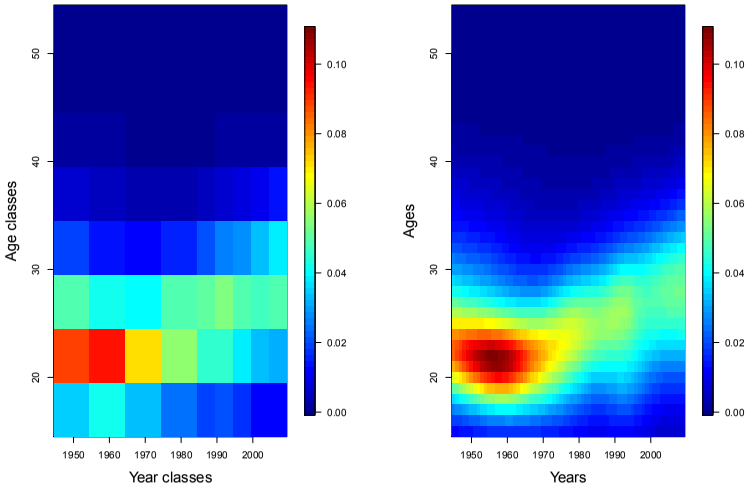
Now we apply the three-dimensional CLMM approach to the previous grouped counts, in order to obtain detailed trends at a fine resolution. In this case we consider the known female population exposure at the fine resolution as $\boldsymbol{e}_{\mathrm{f}}$. To set up the CLMM formulation, we use $\boldsymbol{x}_1 = (15, ..., 54)'$ (for ages), $\boldsymbol{x}_2 = (1945, ..., 2009)'$ (for years), and $\boldsymbol{x}_3 = (1, ..., 4)'$ (for birth orders) as covariates at the fine resolution. We choose 10, 16, and 3 equally-spaced knots for the marginal cubic B-spline bases $\mathbf{B}_d$, respectively, and second order penalties, for $d = 1, 2, 3$. About the composition matrix $\mathbf{C}$ in Eq. (2.46), since we have not aggregated the data by birth order, the marginal composition matrix $\mathbf{C}_3$ is equal to $\mathbf{I}_4$. The dimensions of the other marginal composition matrices $\mathbf{C}_1$ and $\mathbf{C}_2$ are 8×40 and 9×65, respectively.

The right panels of Figure 2.5a, Figure 2.5b, Figure 2.5c, and Figure 2.5d show the resulting CLMM estimates for the fertility rates along ages and years,

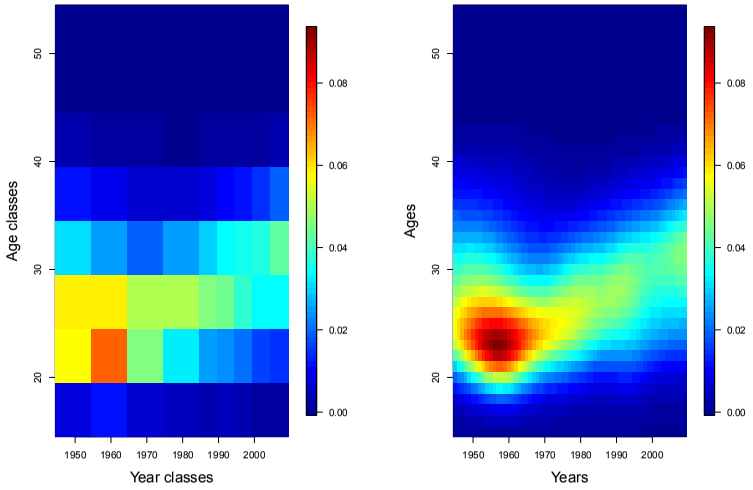
for the 1st, 2nd, 3rd, and 4th birth order, respectively. We observe more detailed insights of the Canadian fertility, delineating clearly lower and higher fertility rates. Figure 2.6a shows the estimated fertility rates using the CLMM approach at age 36, for 1st, 2nd, 3rd, and 4th birth orders. We observe the patterns that the solid curves exhibit follow closely the true distribution of the fertility rates (dot points), except perhaps in the left extremes of some curves where we have less information (recall the aggregation procedure previously done). Figure 2.6b shows the estimated fertility rates using the CLMM approach for year 1990. In this case, we have calculated the true and the estimated total fertility rates, obtaining similar results. Notice here that, in the case of the 1st birth order, the red curve departs slightly from the raw distribution from ages 15 to 27. This can be improved if we incorporate more information in the left part of the curve, i.e., fertility information below 15 years old.

2.4 Summary of the chapter

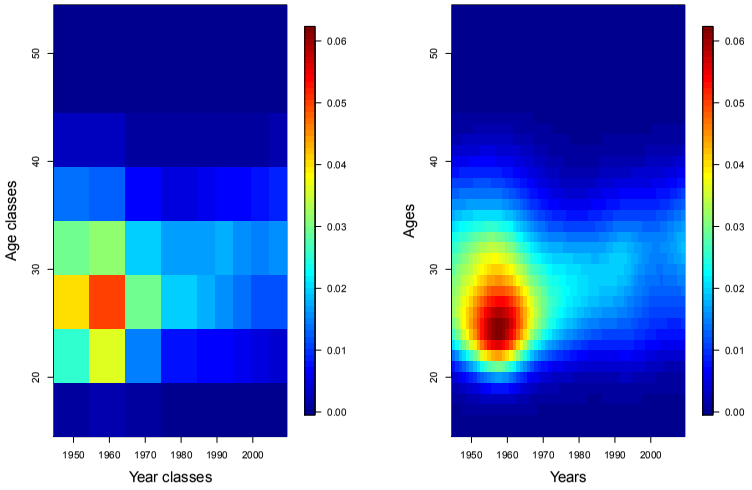
In this chapter we introduced the composite link mixed model approach for grouped count data. It is a generalization of the penalized composite link model (Eilers, 2007), which allows us to incorporate a vector of exposures defined at a fine scale (to analyse rates instead of counts) and more complex structures in terms of random effects. We have presented a parameter estimation procedure for the composite link mixed model, and extended the approach to the multidimensional array case. The latter is achieved by using the Kronecker product operator, thus allowing the use of GLAM algorithms to speed up computations. Several examples were presented in order to illustrate our methodology for the unidimensional and multidimensional array cases.



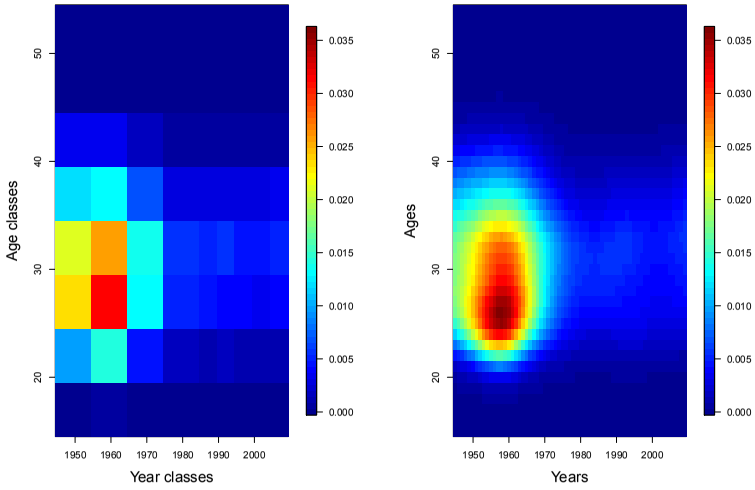
(a) 1st birth



(b) 2nd birth

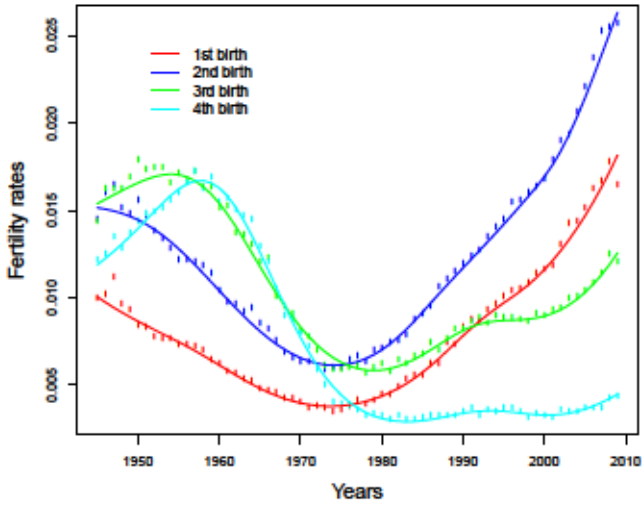


(c) 3rd birth

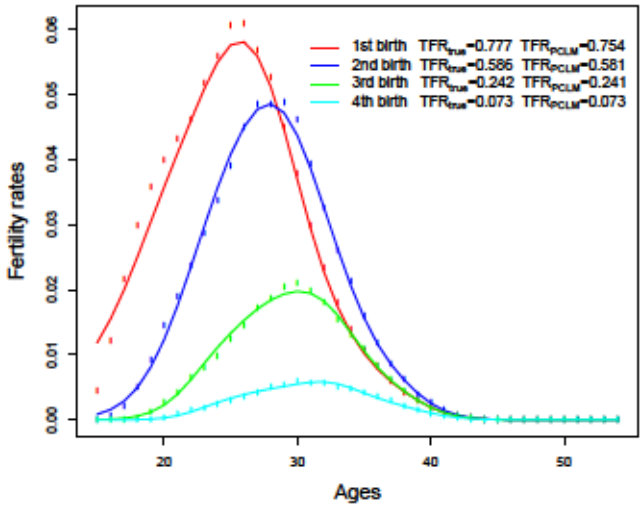


(d) 4th birth

Figure 2.5: Grouped Canadian fertility rates (left side of each panel) for 1st, 2nd, 3rd, and 4th birth orders, and the estimated fertility rates using the three-dimensional CLMM approach (right side of each panel).



(a) Fertility rates at age 36



(b) Fertility rates in 1990

Figure 2.6: True (dot points) and estimated (solid lines) fertility rates (a) at age 36; and (b) for year 1990.

Chapter 3

Estimation of latent spatial trends with the composite link mixed model approach

In this chapter we develop a novel methodology for the analysis of spatially aggregated data, based on the CLMM approach introduced in Chapter 2. The spatial CLMM allows us to create mortality risk maps, from these spatially aggregated data, at a desirable fine resolution. As we have seen in Section 1.2, this fine resolution can be determined by a fine grid, i.e., area-to-point (ATP) case, or by smaller geographical units than the original ones, i.e., the area-to-area (ATA) case. The spatial CLMM can handle this two types of spatial disaggregation in an elegant way. Moreover, in this chapter we present a methodology to take into account the problem of overdispersion in count data. Part of the results given here are published in Ayma et al. (2016).

This chapter is organized as follows. In Section 3.1 we present the extension of the CLMM approach to develop a new methodology for the ATP and ATA cases. In Section 3.2 we present a methodology to deal with the problem of overdispersion in count data, by incorporating individual random effects at the aggregated scale. In Section 3.3 we illustrate the spatial CLMM approach for the ATP case, where our methodology is compared with the ATP Poisson kriging of Goovaerts (2006). Here, we use two datasets related with female deaths by lung cancer in Indiana, USA, and male lip cancer incidence in Scotland counties. In

Section 3.4 we illustrate our methodology for the ATA case, where we use a female death dataset by cardiovascular diseases in the Community of Madrid. Also, we illustrate here the inclusion of explanatory variables, defined at a fine scale, into the CLMM formulation. Finally, in Section 3.5 we give a summary of this chapter.

3.1 The spatial composite link mixed model approach

In order to present the extension of the CLMM approach given in Section 2.2 to the spatial case, we briefly review the P-spline methodology and its mixed model formulation to the spatial case.

3.1.1 P-spline methodology for spatial data

As we have seen in Chapter 2, the P-spline methodology is based on the use of a regression basis and a penalty matrix that controls the amount of smoothness of the fitted curve. In a two-dimensional context, where the spatial case is included, there exist several approaches for the election of the basis and the penalty (see, for example, Ruppert et al., 2003; Kammann and Wand, 2003; Wood, 2006b). However, most of them assume the same amount of smoothness in both directions (i.e., isotropy property) and/or the selection of the knots (to construct the regression basis) is not easy to implement. Here we follow the proposal of Lee and Durbán (2009) and use a tensor product of B-spline bases with equally-spaced knots.

For simplicity, assume we have normally distributed spatial data $(x_{1j}, x_{2j}, y_j)'$, $j = 1, ..., m$, where $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are the geographical coordinates (longitude and latitude, respectively) and $\boldsymbol{y}$ is the response variable. Then, a smooth model for the data is given by:

$$\boldsymbol{y} = f(\boldsymbol{x}_1, \boldsymbol{x}_2) + \boldsymbol{\epsilon} = \mathbf{B}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \backsim \mathcal{N}(\mathbf{0}, \sigma^2_{\epsilon}\mathbf{I}), \tag{3.1}$$

where $\mathbf{B}$ is a regression basis constructed from $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, and $\boldsymbol{\theta}$ is the vector of regression coefficients. To achieve smoothness, a penalty matrix $\mathbf{P}$ is introduced by modifying the corresponding sum of squares (see Eq. (2.9)). Following the proposal of Lee and Durbán (2009), the regression basis $\mathbf{B}$ is constructed as the

Box-product or 'row-wise' Kronecker product (Eilers et al., 2006) of two marginal B-spline bases $\mathbf{B}_1 = \mathbf{B}(\boldsymbol{x}_1)$ and $\mathbf{B}_2 = \mathbf{B}(\boldsymbol{x}_2)$ of dimensions $m \times c_1$ and $m \times c_2$, respectively:

$$\mathbf{B} = \mathbf{B}_2 \square \mathbf{B}_1 = (\mathbf{B}_2 \otimes \mathbf{1}'_{c_1}) \odot (\mathbf{1}'_{c_2} \otimes \mathbf{B}_1), \tag{3.2}$$

where $\odot$ represents the Hadamard or 'element-wise' product. The matrix operator $\square$ in Eq. (3.2) receives the name of 'row-wise' Kronecker product, because the i-th row of $\mathbf{B}_2 \square \mathbf{B}_1$ in Eq. (3.2) is the Kronecker product of the i-th rows of $\mathbf{B}_2$ and $\mathbf{B}_1$. With this new regression basis, the geographical coordinates are not subject to be on a regular grid.

Regarding the penalty matrix $\mathbf{P}$ for the spatial case, it is the same as in the two-dimensional array case given in Eq. (2.43). This is because the penalty matrix acts on the regression coefficients $\boldsymbol{\theta}$ (which can be always arranged in array form) and does not depend on the data structure. Moreover, since it is an anisotropic penalty (i.e., $\lambda_1 \neq \lambda_2$), it is suitable for cases when different degrees of smoothing for each dimension are needed, or when covariates are measured in different scales.

Within the P-spline framework, the fitted values of the model given in Eq. (3.1) are obtained as in Eq. (2.10), with $\mathbf{B}$ and $\mathbf{P}$ defined in Eq. (3.2) and Eq. (2.43), respectively. Optimal values for the smoothing parameters λ_1 and λ_2 can be obtained by minimizing an information criterion such as AIC or BIC (as in the multidimensional array case).

The modelling of non-normal data, as in the case of spatial count data, is achieved by extending the P-spline methodology to the GLM setting. Moreover, it can be extended to the GLMM framework, where the spatial B-spline basis and the two-dimensional discrete penalty are reformulated into a mixed model setting (see Lee and Durbán, 2009, for more details). The later is briefly described below.

3.1.2 Spatial mixed model formulation of P-splines

As in the unidimensional and multidimensional array cases, it is possible to reparameterize the spatial regression basis $\mathbf{B}$ in Eq. (3.2) into the mixed model framework (i.e., $\mathbf{B} \rightarrow [\mathbf{X} : \mathbf{Z}]$). This was shown by Lee (2010), where the orthogonal transformation matrix $\mathbf{T}$ in Eq. (2.48) is used to achieve the desire reparameteri-

zation. The resulting mixed model matrices for spatial data are:

$$\begin{aligned} \mathbf{X} &= \mathbf{X}_2 \square \mathbf{X}_1, \\ \mathbf{Z} &= [\mathbf{Z}_2 \square \mathbf{X}_1 : \mathbf{X}_2 \square \mathbf{Z}_1 : \mathbf{Z}_2 \square \mathbf{Z}_1], \end{aligned} \tag{3.3}$$

where $\mathbf{X}_d = \mathbf{B}_d \mathbf{U}_{dn}$ and $\mathbf{Z}_d = \mathbf{B}_d \mathbf{U}_{ds}$, for $d = 1, 2$. Notice that the structure of the mixed model matrices in Eq. (3.3) is similar to those given in Eq. (2.49) (for the array case), but with $\square$ operators instead of $\otimes$. This reparameterization is not straightforward as it may seem and special matrix algebra results are needed.

On the other hand, and since we are using the penalty matrix $\mathbf{P}$ in Eq. (2.43), the mixed model penalty matrix for the spatial case is the same as in the two-dimensional array case, with the mixed model block-diagonal matrix $\mathbf{F}$ defined in (2.50).

3.1.3 Spatial composite link mixed models

Considering the spatial mixed model reformulation given above, we can generalize the CLMM approach introduced in Chapter 2 to the spatial case.

Suppose that we observe a vector of aggregated counts $\boldsymbol{y}$, assumed Poisson distributed with mean vector $\boldsymbol{\mu}$, which are available over n non-overlapping geographical units v_i, $i = 1, ..., n$. Let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be the geographical coordinates (longitude and latitude, respectively) of length m that define the desirable fine spatial resolution. Then, the spatial CLMM is given as:

$$\boldsymbol{\mu} = \mathbf{C}_s \boldsymbol{\gamma} = \mathbf{C}_s(\boldsymbol{e}_f * \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha})), \text{ with } \boldsymbol{\alpha} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{G}(\lambda_1, \lambda_2)\right), \tag{3.4}$$

where now $\mathbf{X}$ and $\mathbf{Z}$ are the spatial mixed model matrices defined in Eq. (3.3), with the inverse of the covariance matrix $\mathbf{G}$ for the random effects coefficients $\boldsymbol{\alpha}$ given as in Eq. (2.50). The vector $\boldsymbol{e}_f$ is considered now as the vector of exposures at the fine spatial resolution, where geographical units $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are defined.

The new spatial composition matrix $\mathbf{C}_s$ is fixed (as usual) and its structure depends on the relationship between the coarse and the fine spatial resolutions (defined by the geographical units v_i, and the fine-scale geographical coordinates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively). Here we illustrate two possible cases: the first one where the fine scale resolution is given by a fine grid over the study region (which is

formed by the units v_i), and the second one where the fine resolution is given by small geographical units that are contained in the coarser units v_i. These two situations are called the area-to-point (ATP) and area-to-area (ATA) cases, respectively. Thus, we consider the coordinates $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ in the ATP case as the centroids of the grid cells, whereas in the ATA case we consider them as the centroids of the smaller units. These two cases were previously illustrated in Chapter 1. Therefore, the elements of the spatial composition matrix $\mathbf{C}_\mathrm{s}$ for the ATP and ATA cases are given by:

$$c_{ij} = \begin{cases} 1 & \text{if } (x_{1j}, x_{2j}) \text{ belongs to unit } v_i \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

for $i = 1, ..., n$, and $j = 1, ..., m$, where (x_{1j}, x_{2j}) are the spatial coordinates for the fine scale resolution.

Since our goal is to analyse rates, the vector $\boldsymbol{e}_\mathrm{f}$ has to be known in advance; otherwise, it has to be estimated. If the vector of exposures is only available at the aggregated level, a naive approach to estimate $\boldsymbol{e}_\mathrm{f}$ is to assume that these aggregated exposures are evenly distributed throughout the fine resolution. Another possibility is to apply the CLMM approach to the aggregated vector of exposures to obtain estimates for $\boldsymbol{e}_\mathrm{f}$.

The parameter estimation of the spatial CLMM given in Eq. (3.4) can be carried out by using the procedure given in Section 2.2. From (3.5), we see that the spatial composition matrix $\mathbf{C}_\mathrm{s}$ is sparse (i.e., most of its elements are zero) and, thus, sparse methods can be used in order to speed up computations (see, for example, the packages of Bates and Maechler, 2015, and of Koenker and Ng, 2016).

Before illustrating the spatial CLMM methodology for the ATA and ATP cases, we will see in the next section how our approach can handle the overdispersion problem, which frequently appears when we analyse count data.

3.2 Handling overdispersion with CLMMs

The CLMM approach is a useful tool for modelling aggregated or grouped counts. However, it is assumed the counts for the groups follow Poisson distributions.

When this is not the case, because of overdispersion (i.e., the presence of extra Poisson variation due to an unobserved heterogeneity), underestimation of the variability of estimates may occur. As a solution for the overdispersion problem, we propose to introduce individual random effects for the logarithms of the expected values, one for each group count. This can be viewed as an adaptation of the PRIDE ('penalized regression with individual deviance effects') approach given by Perperoglou and Eilers (2010) and Lee and Durbán (2009). Here, we develop this idea under the CLMM framework; thus we will refer to this approach as CLMM-P.

Consider $\boldsymbol{\phi} = \mathbf{C}\boldsymbol{\gamma}$, where $\mathbf{C}$ is the composition matrix and $\boldsymbol{\gamma}$ is the vector of latent expectations at the fine resolution, with $\boldsymbol{\gamma} = \boldsymbol{e}_{\mathrm{f}} * \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha})$. We can generalize the CLMM formulation by assuming that the aggregated counts are now Poisson distributed with mean vector:

$$\boldsymbol{\mu} = \exp(\log(\boldsymbol{\phi}) + \boldsymbol{\delta}),\ \boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}),\ \boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \kappa^{-1}\mathbf{I}_n), \tag{3.6}$$

where κ is the dispersion parameter associated with the individual random effects $\boldsymbol{\delta}$. These random effects (defined at the aggregated scale) provides a device to absorb the overdispersion that causes the extra-variability. Thus, in the model given by Eq. (3.6), we are simultaneously dealing with parameters at aggregated and at a finer scale.

Considering the penalized log-likelihood:

$$\ell_{\mathrm{P}}^{*} = \log\left(f(\boldsymbol{y}|\boldsymbol{\alpha}, \boldsymbol{\delta})\right) - \frac{1}{2}\boldsymbol{\alpha}'\mathbf{G}^{-1}\boldsymbol{\alpha} - \frac{1}{2}\kappa\boldsymbol{\delta}'\boldsymbol{\delta},$$

where $f(\boldsymbol{y}|\boldsymbol{\alpha}, \boldsymbol{\delta})$ denotes the joint density distribution of $\boldsymbol{y}$ in the CLMM-P context, and using the PQL approach for the estimation of the parameters $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\delta}$ in Eq. (3.6), we obtain the following system of equations:

$$\begin{bmatrix} \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{X}} & \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}} & \breve{\mathbf{X}}'\mathbf{W} \\ \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{X}} & \mathbf{G}^{-1} + \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}} & \breve{\mathbf{Z}}'\mathbf{W} \\ \mathbf{W}\breve{\mathbf{X}} & \mathbf{W}\breve{\mathbf{Z}} & \kappa\mathbf{I}_n + \mathbf{W} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \\ \boldsymbol{\delta} \end{bmatrix} = \begin{bmatrix} \breve{\mathbf{X}}'\mathbf{W}\boldsymbol{z} \\ \breve{\mathbf{Z}}'\mathbf{W}\boldsymbol{z} \\ \mathbf{W}\boldsymbol{z} \end{bmatrix}, \tag{3.7}$$

where now the 'working' matrices are defined as $\breve{\mathbf{X}} = \boldsymbol{\Phi}^{-1}\mathbf{C}\boldsymbol{\Gamma}\mathbf{X}$ and $\breve{\mathbf{Z}} = \boldsymbol{\Phi}^{-1}\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z}$, with $\boldsymbol{\Phi} = \mathrm{diag}(\boldsymbol{\phi})$ and $\boldsymbol{\Gamma} = \mathrm{diag}(\boldsymbol{\gamma})$. In this case, the matrix of weights and the working vector are $\mathbf{W} = \mathrm{diag}(\boldsymbol{\mu})$, with $\boldsymbol{\mu}$ defined as in Eq. (3.6), and $\boldsymbol{z} =$

$\breve{\mathbf{X}}\boldsymbol{\beta} + \breve{\mathbf{Z}}\boldsymbol{\alpha} + \boldsymbol{\delta} + \mathbf{W}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$, respectively.

It is possible to reduce the large system of equations given in Eq. (3.7) by defining $\boldsymbol{\delta}$ as:

$$\boldsymbol{\delta} = (\mathbf{W} + \kappa\mathbf{I}_n)^{-1}\mathbf{W}(\boldsymbol{z} - \breve{\mathbf{X}}\boldsymbol{\beta} - \breve{\mathbf{Z}}\boldsymbol{\alpha}). \tag{3.8}$$

Thus, if we define:

$$\mathbf{W}^* = \kappa(\mathbf{W} + \kappa\mathbf{I}_n)^{-1}\mathbf{W},$$

we have that $\kappa\boldsymbol{\delta} = \mathbf{W}^*(\boldsymbol{z} - \breve{\mathbf{X}}\boldsymbol{\beta} - \breve{\mathbf{Z}}\boldsymbol{\alpha})$. Using this result in Eq. (3.7), we obtain:

$$\begin{bmatrix} \breve{\mathbf{X}}'\mathbf{W}^*\breve{\mathbf{X}} & \breve{\mathbf{X}}'\mathbf{W}^*\breve{\mathbf{Z}} \\ \breve{\mathbf{Z}}'\mathbf{W}^*\breve{\mathbf{X}} & \mathbf{G}^{-1} + \breve{\mathbf{Z}}'\mathbf{W}^*\breve{\mathbf{Z}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} \breve{\mathbf{X}}'\mathbf{W}^*\boldsymbol{z} \\ \breve{\mathbf{Z}}'\mathbf{W}^*\boldsymbol{z} \end{bmatrix}.$$

This leads to the same system of equations of the Poisson CLMM without overdispersion (see Eq. (2.32)), but changing the matrix of weights to $\mathbf{W}^*$ and the addition of the vector $\boldsymbol{\delta}$ to the working vector. Therefore, the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are estimated as in Eq. (2.33), with $\mathbf{V} = \mathbf{W}^{*-1} + \breve{\mathbf{Z}}\mathbf{G}\breve{\mathbf{Z}}'$, and $\boldsymbol{\delta}$ is estimated using Eq. (3.8). Then, conditioning on these estimates, the smoothing parameters (λ_1 and λ_2) and the dispersion parameter (κ) are estimated by the approximate REML in (2.37).

To compute the effective dimension of the model given in Eq. (3.6), the hat matrix $\mathbf{H}^*$ for the CLMM-P is given by:

$$\mathbf{H}^* = [\breve{\mathbf{X}} : \breve{\mathbf{Z}} : \mathbf{I}_n] \begin{bmatrix} \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{X}} & \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}} & \breve{\mathbf{X}}'\mathbf{W} \\ \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{X}} & \mathbf{G}^{-1} + \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}} & \breve{\mathbf{Z}}'\mathbf{W} \\ \mathbf{W}\breve{\mathbf{X}} & \mathbf{W}\breve{\mathbf{Z}} & \kappa\mathbf{I}_n + \mathbf{W} \end{bmatrix}^{-1} \begin{bmatrix} \breve{\mathbf{X}}'\mathbf{W} \\ \breve{\mathbf{Z}}'\mathbf{W} \\ \mathbf{W} \end{bmatrix}.$$

3.3 CLMM application to area-to-point case

In this section we apply our methodology to two real datasets. We use the first dataset to illustrate the spatial CLMM approach for the ATP case. With the second dataset, we illustrate how the CLMM-P approach can handle the problem of overdispersion, often present in count data. For parameter estimation, we follow the methodology described in Section 2.2. Using both datasets, we also compare our methodology with the ATP Poisson kriging of Goovaerts (2006). Hereafter we refer to this approach as PK. We start first with a briefly description of the PK

approach for the ATP case.

3.3.1 ATP Poisson kriging

The ATP Poisson kriging approach was developed by Goovaerts (2006) and is implemented in the geostatistical software SpaceStat 4.0 (http://www.biomedware.com/). Given the fine grid point $\boldsymbol{u}_j = (x_{1j}, x_{2j})$, for $j = 1, ..., m$, within a geographical unit v_δ, the PK estimator is obtained as a linear combination of the kernel rate $r(v_\delta)$ and the rates observed in $(K-1)$ neighbouring units:

$$\widehat{r}_{\text{PK}}(\boldsymbol{u}_j) = \sum_{i=1}^{K} \lambda_i(\boldsymbol{u}_j) r(v_i), \tag{3.9}$$

where $\lambda_i(\boldsymbol{u}_j)$ is the kriging weight assigned to the rate $r(v_i)$ when estimating the risk at $\boldsymbol{u}_j$. The PK variance associated to the estimator given in Eq. (3.9) is computed as:

$$\sigma^2_{\text{PK}}(\boldsymbol{u}_j) = \bar{C}_R(0) - \sum_{i=1}^{K} \lambda_i(\boldsymbol{u}_j) \bar{C}_R(v_i, \boldsymbol{u}_j) - \mu(\boldsymbol{u}_j). \tag{3.10}$$

In Eq. (3.10), $\bar{C}_R(v_i, \boldsymbol{u}_j)$ are area-to-point covariances that are approximated as:

$$\bar{C}_R(v_i, \boldsymbol{u}_j) = \left(\sum_{j'=1}^{P_i} w_{jj'} \right)^{-1} \times \sum_{j'=1}^{P_i} w_{jj'} C(\boldsymbol{u}_j, \boldsymbol{u}_{j'}),$$

where P_i is the number of points used to discretize the unit v_i and the weights $w_{j'j}$ are computed as $w_{jj'} = n(\boldsymbol{u}_j) \times n(\boldsymbol{u}_{j'})$, with $n(\boldsymbol{u}_j)$ being the population size within the square cell centred on $\boldsymbol{u}_j$. Then, the K kriging weights and the Lagrange parameter $\mu(\boldsymbol{u}_j)$ are computed by solving the following system of linear equations:

$$\sum_{k=1}^{K} \lambda_k(\boldsymbol{u}_j) \left(\bar{C}_R(v_i, v_k) + \delta_{ik} \frac{m^*}{n(v_i)} \right) + \mu(\boldsymbol{u}_j) = \bar{C}_R(v_i, \boldsymbol{u}_j),\ i = 1, ..., K$$

$$\sum_{k=1}^{K} \lambda_k(\boldsymbol{u}_j) = 1$$

where $\delta_{ik} = 1$ if $i = k$ and 0 otherwise, m^* is the population-weighted mean of the n observed rates, $n(v_i)$ is the size of the population-at-risk in unit v_i, and $\bar{C}_R(v_i, v_k)$ are area-to-area covariances that are approximated in a similar fashion as the area-to-point covariances (see Eq. 8 in Goovaerts, 2006).

To solve the previous ATP kriging system, the point-support covariance of the risk $C(\boldsymbol{h})$, or equivalently a point-support semivariogram $\gamma(\boldsymbol{h})$ has to be known in advance. Since only aggregated data are available, this function cannot be estimated directly from the observed rates. Goovaerts (2008) developed an iterative procedure to conduct the derivation of $\gamma(\boldsymbol{h})$ from the 'regularized' experimental semivariogram computed from areal data (i.e., 'deconvolution' process), in presence of irregular geographical units and heterogeneous population distribution. See Goovaerts (2008) for a detailed presentation of the deconvolution procedure and demonstration of its performances in simulation studies.

Once we have briefly presented the PK approach, we proceed to apply the CLMM, CLMM-P, and PK approaches to two real datasets: female deaths by lung cancer in Indiana, USA, and male lip cancer incidence in Scotland counties.

3.3.2 Application 1: Lung cancer dataset

The lung cancer dataset comes from the Atlas of Cancer Mortality in the United States (Pickle et al., 1999), and can be downloaded from http://ratecalc.cancer.gov. This dataset has been previously analysed by Goovaerts (2006), and it contains the number of white female deaths by lung cancer and the corresponding age-adjusted mortality rates (per 100000 person-years), recorded over the period 1970-1994 in the state of Indiana at county level (92 counties in total). The population-at-risk in each county can be estimated with the formula:

$$\frac{\text{Total number of deaths (1970-1994)}}{\text{Age-adjusted mortality rate (1970-1994)}} \times 10^5 \ .$$

Goovaerts (2006) imposed a 55×94 grid (with grid cells of 25 km^2) over the map of Indiana, leading to 3751 grid points that fall inside the map (see Figure 3.1a). Next, he allocated the previous county-level population estimates to this fine grid, according to the 2000 census block level data. Figure 3.1b shows the spatial distribution of the population-at-risk on the fine grid, which reflects the heterogeneous

repartition of population in Indiana. These high-resolution population estimates were kindly provided by Dr. Pierre Goovaerts (BioMedware Inc., MI, USA) and we will use them in subsequent analysis.

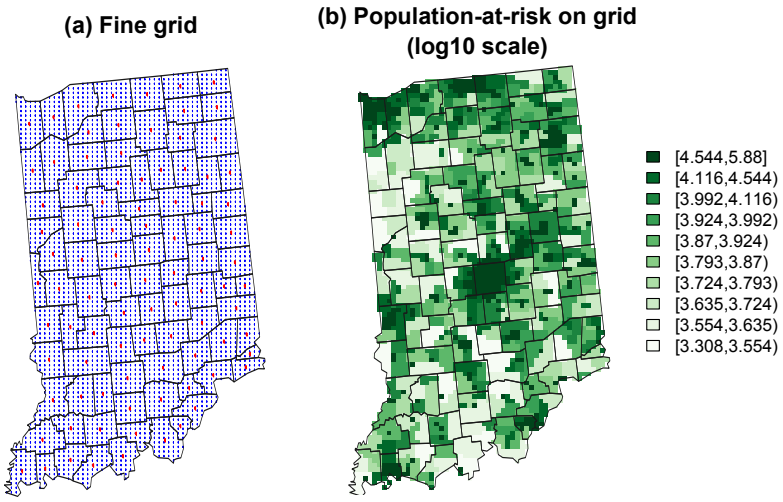


Figure 3.1: The left map shows the fine grid obtained by imposing a 55×94 grid over the map of Indiana, leading to 3751 grid points (blue points). The right map shows the spatial distribution of the population-at-risk on this fine grid (on a log10 scale).

Figure 3.2a shows the spatial distribution of age-adjusted mortality rates (per 100000 person-years) for lung cancer in Indiana. We use a yellow-red color scheme for data visualization, where the class boundaries correspond to the deciles of the original rates. Rates higher than the median tend to be more red as they depart from it, while lower rates tend to be more yellow. Since the sizes of the counties in Indiana are relatively similar, it is easy to identify areas of excess in this region. The highest rates are reported for the counties of Clark (30.637), Johnson (30.726), and Marion (31.624), which is the most populated county in Indiana.

To reduce the noise present in lung cancer mortality rates, we first apply the PGLMM approach (Lee and Durbán, 2009) with the spatial coordinates of the county centroids as covariates, second order penalties, and 22 equally-spaced knots for each marginal cubic B-spline basis. Figure 3.2b shows the resulting smoothed mortality risk, with range varying from 13.302 to 31.624. The maximum rates after smoothing are still located in counties with the highest lung cancer rates. This situation was also pointed out by Goovaerts (2006), when he analysed these data

(at county level) with different kriging methods. For this dataset, if we increase the number of knots in the PGLMM approach, we will obtain a similar spatial mortality pattern to that shown in Figure 3.2b.

Now we apply the CLMM approach on this dataset to obtain a continuous mortality risk map. To do that, let us consider the number of white female deaths by lung cancer per county as the vector of aggregated counts ($\boldsymbol{y}$), and the population-at-risk on the fine grid of 25 km^2 cells (displayed in Figure 3.1b) as the vector of exposures at fine resolution ($\boldsymbol{e}_{\mathrm{f}}$). To set up the CLMM formulation, we use the spatial coordinates of the grid points (see Figure 3.1a) as covariates at fine resolution, second order penalties, and 22 equally-spaced knots for each marginal cubic B-spline basis. Then, we can construct the spatial composition matrix as is described in Eq. (3.5). Figure 3.2c shows the resulting CLMM mortality risk, which is calculated as $\widehat{r}_{\mathrm{CLMM}} = 10^5 \times \exp(\mathbf{X}\widetilde{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\boldsymbol{\alpha}})$. This isopleth map gives a more detailed impression of the mortality distribution, where areas of lower and higher mortality risks are clearly delineated on the map of Indiana. Higher risk estimates are still observed in the counties of Clark, Johnson, and Marion, while lower risk estimates are more concentrated in some south-western and north-eastern counties of Indiana.

To compare our proposal with other existing methods, we apply the PK approach of Goovaerts (2006) to this dataset. Figure 3.2d shows the resulting PK mortality risk, using the software indicated above, together with the indications given in Goovaerts (2006) for the estimation of this continuous surface. The PK approach provides a similar spatial pattern to the CLMM approach, with some discrepancies in the north and south-east of the central counties. We should note that the application of both approaches to this dataset produces some risk estimates at fine scale that exceed the maximum raw lung mortality rate (31.795). For example, the maximum risk estimates for the CLMM and PK are 34.067 and 33.896, respectively.

Figure 3.3 shows the standard error maps associated with the mortality risk maps given at the bottom of Figure 3.2. The PK standard errors are calculated as the square root of the PK variances (see Eq. (3.10)). Most of the CLMM standard errors are lower than those obtained with the PK, through Indiana counties, showing that CLMM reduces the uncertainty.

To compare the aggregations resulting from the CLMM and PK approaches,

we can compute the corresponding AIC using the estimated means $\hat{\boldsymbol{\mu}}_{\text{CLMM}}$ and $\hat{\boldsymbol{\mu}}_{\text{PK}}$, respectively. The first one is calculated as in Eq. (3.4), while the elements of the second are obtained from Goovaerts (2006, Eq. 15) as:

$$\hat{\mu}_{\text{PK}}(v_i) = 10^{-5} \times e(v_i)\hat{r}_{\text{PK}}(v_i) = 10^{-5} \times \sum_{j=1}^{P_i} e_{\text{f}}(\boldsymbol{u}_j)\hat{r}_{\text{PK}}(\boldsymbol{u}_j), \tag{3.11}$$

where P_i denotes the number of grid points used to discretize the county v_i, and $e(v_i) = \sum_{j=1}^{P_i} e_{\text{f}}(\boldsymbol{u}_j)$, for $i = 1, ..., n$. The resulting AIC for the CLMM and PK (at county level) are 163.565 and 237.394, respectively. Therefore, the CLMM approach is preferable in this case.

In order to assess the prediction performance among the mentioned approaches, we have carried out a simulation study in Section 3.3.4.

3.3.3 Application 2: Scottish lip cancer dataset

The Scottish lip cancer dataset (Clayton and Kaldor, 1987) has been widely analysed in the literature, and is typically used to illustrate the problem of overdispersion for areal data. Here, the data are recorded over 56 counties (see Figure 1.2a), whose sizes and shapes vary considerably. In this subsection, we apply the CLMM-P approach (developed in Section 3.2) on this dataset to obtain a continuous surface that take into account the overdispersion present in count data.

This dataset consists of the observed ($\boldsymbol{y}$) and expected ($\boldsymbol{e}$) number of male cases of lip cancer, recorded in 56 counties in Scotland over the period 1975-1980. Figure 3.4a shows the spatial distribution of the Standardized Mortality Rates (SMRs) on a logarithmic scale for lip cancer incidence, which are obtained as:

$$\log(\text{SMR})_i = \log\left(\frac{y_i}{e_i}\right), \text{ for } i = 1, ..., 56.$$

We see that most of the higher raw log(SMRs) are located in the north of Scotland; specifically in the counties of Caithness, Ross and Cromarty, Skye and Lochalsh, and Banff and Buchan. For graphical simplicity, notice that the northest Scottish islands in Figure 1.2a were moved in Figure 3.4a.

In order to apply the CLMM approach, we impose a 120×120 fine grid over the map of Scotland, leading to 3855 grid points that fall inside the map. Since

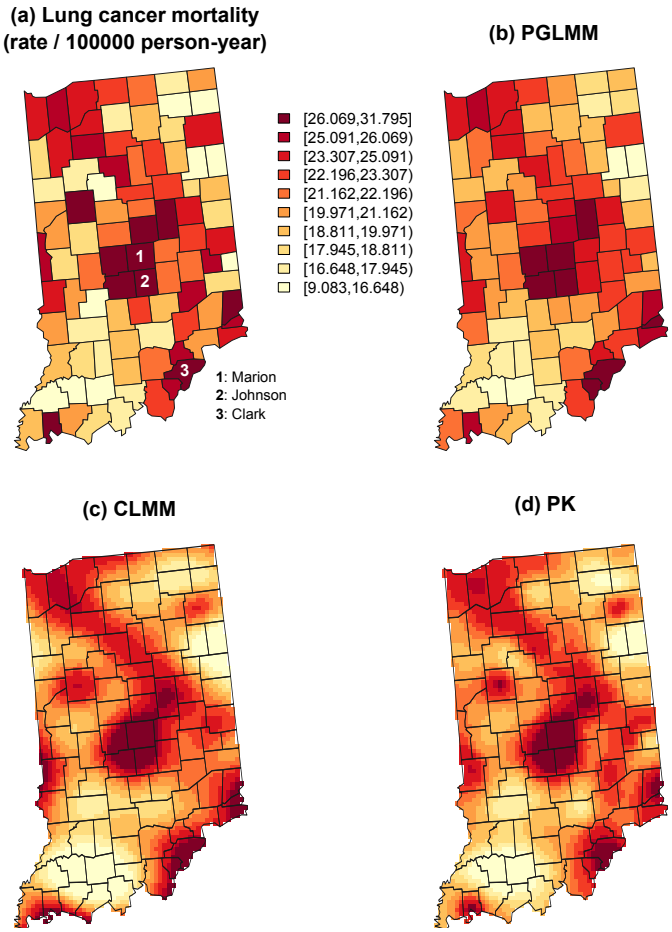


Figure 3.2: Map of lung cancer mortality rates in Indiana, and the risk estimated by different approaches. The top-left map displays the age-adjusted mortality rates per 100000 person-years recorded over the period 1970-1994, and the top-right map shows the smoothed mortality risks resulting from the PGLMM approach. The bottom maps show the smoothed mortality risks estimated using the CLMM (bottom-left) and PK (bottom-right) approaches. The color legend applies to all maps; the class boundaries correspond to the deciles of the original rates.

the vector of exposures is unavailable at this fine scale, we estimate it using the naive approach described in Section 3.1. We denote this vector as $\hat{\boldsymbol{e}}_{\text{naive}}$. To set up the CLMM formulation, we use 25 equally-spaced knots for each marginal cubic B-spline basis and second order penalties. Then, the corresponding spatial composition matrix is constructed as is described in Eq. (3.5). Figure 3.4b shows

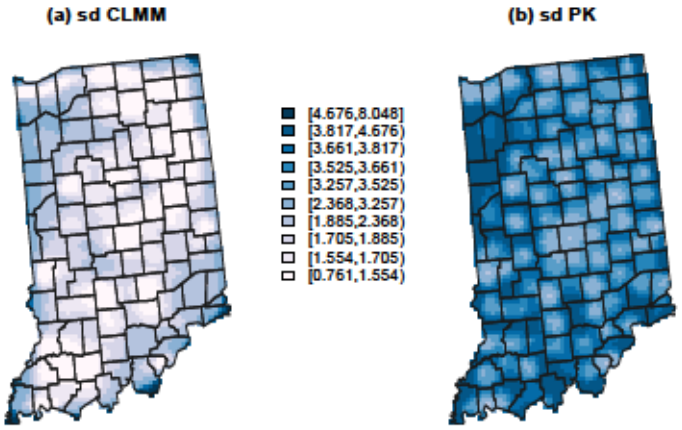


Figure 3.3: Standard error maps for lung cancer mortality risk in Indiana, estimated by (a) CLMM and (b) PK approaches.

the resulting CLMM estimates for the log(SMR) at the selected fine grid (that is, $\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\boldsymbol{\alpha}}$). From Figure 3.4b, we observe there exist an increasing trend from the more central counties to the ones of the coast, and also from south to north. Moreover, using the previous point estimates, we can obtain a smooth trend for the log(SMR) at county level as follows:

$$\log(\text{SMR})_{\text{CLMM}} = \log\left(\frac{\hat{\boldsymbol{\mu}}_{\text{CLMM}}}{\boldsymbol{e}}\right), \tag{3.12}$$

where $\hat{\boldsymbol{\mu}}_{\text{CLMM}}$ is obtained as in Eq. (3.4), with $\boldsymbol{e}_{\text{f}} = \hat{\boldsymbol{e}}_{\text{naive}}$. Figure 3.4c shows these estimates for the log(SMR) at county level.

Now we apply the CLMM-P approach to this dataset. For that we use the same settings as in the CLMM approach. Figure 3.4d shows the resulting CLMM-P estimates for the log(SMR) at the selected fine grid, where we include the estimated individual random effects, $\hat{\boldsymbol{\delta}}$, at the fine scale to take into account the overdispersion. This is done by adding the term $\mathbf{C}^{-}\hat{\boldsymbol{\delta}}$ to the estimated spatial trend (that is, $\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\boldsymbol{\alpha}} + \mathbf{C}^{-}\hat{\boldsymbol{\delta}}$), where $\mathbf{C}^{-}$ denotes the Moore-Penrose inverse of $\mathbf{C}$. This matrix can be easily computed as $\mathbf{C}^{-} = (\mathbf{R}^{-1}\mathbf{C})'$, where $\mathbf{R}$ is a diagonal matrix whose elements are the sums of the rows of $\mathbf{C}$. The map displayed in Figure 3.4d presents some differences with respect to the map obtained with the

CLMM approach, especially in the north of Scotland. Similarly to what we did before, we can obtain a smooth trend for the log(SMR) at county level, from the CLMM-P estimates, as:

$$\log(\text{SMR})_{\text{CLMM-P}} = \log\left(\frac{\hat{\boldsymbol{\mu}}_{\text{CLMM-P}}}{\boldsymbol{e}}\right), \tag{3.13}$$

where $\hat{\boldsymbol{\mu}}_{\text{CLMM-P}}$ is obtained as in Eq. (3.6), with $\boldsymbol{e}_{\text{f}} = \hat{\boldsymbol{e}}_{\text{naive}}$. These estimates for log(SMR) at county level are displayed in Figure 3.4e.

To compare our proposal with other existing methods, we apply the PK approach to this dataset. Figure 3.4f shows the resulting PK estimates for the log(SMR) at the selected fine grid, which is similar to that produced by the CLMM approach. Also, we can obtain a smooth trend for the log(SMR) at county level, from the PK estimates, as:

$$\log(\text{SMR})_{\text{PK}} = \log\left(\frac{\hat{\boldsymbol{\mu}}_{\text{PK}}}{\boldsymbol{e}}\right), \tag{3.14}$$

where $\hat{\boldsymbol{\mu}}_{\text{PK}}$ is obtained as in Eq. (3.11). These estimates for log(SMR) at county level are displayed in Figure 3.4g.

Figure 3.5 shows the standard error maps associated with the middle row of Figure 3.4. In this case, we observe that higher errors are located in the islands of the north and north-west of Scotland. In these parts, the errors associated to the CLMM and CLMM-P approaches are greater than those associated with the PK approach. The higher standard errors in CLMM and CLMM-P approaches might be due to the presence of the islands where there is a discontinuity in the boundaries (the tensor product smooth tends to interpolate the sea where no data are available leading to larger standard errors), while PK model implemented in Spacestat 4.0 uses an areal deconvolution process with the definition of a spatial weight matrix with a minimum distance to ensure that all units will be connected with at least one other unit (Jacquez et al., 2014). Some advances in spline smoothing can be studied to include special penalties to account for smoothing in complex and irregular domains (see Ramsay, 2002; Wood et al., 2008).

In order to compare the aggregations resulting from the CLMM, CLMM-P and PK approaches, we can compute the AIC using the estimated means $\hat{\boldsymbol{\mu}}_{\text{CLMM}}$, $\hat{\boldsymbol{\mu}}_{\text{CLMM-P}}$ and $\hat{\boldsymbol{\mu}}_{\text{PK}}$ already calculated in Eq. (3.12)-(3.14), respectively. The result-

ing AIC for the CLMM, CLMM-P and PK (at county level) are 110.8, 89.8, and 186.7, respectively, showing that the CLMM-P is more appropriate in presence of overdispersion.

3.3.4 Simulation study

In this section we perform a simulation study to compare the prediction performance of the CLMM approach with the ATP Poisson kriging (PK) of Goovaerts (2006). For that, we use the lung cancer dataset described in Section 3.3.2. The simulation study was conducted as follows:

1) The continuous mortality surface obtained with the PK approach was considered here as the true underlying mortality trend over the fine grid of 25 km^2 cells in Indiana. We denoted these mortality rates as $r(\boldsymbol{u}_j)$, $j = 1, ..., 3751$, where $\boldsymbol{u}_j$ represent the coordinates of the fine grid points.

2) These quantities and the population-at-risk over each 25 km^2 cell of the fine grid (denoted as $e(\boldsymbol{u}_j)$) were used to calculate the mortality rate for each county v_i, $i = 1, ..., 92$:

$$r(v_i) = \frac{1}{e(v_i)} \sum_{j=1}^{P_i} e(\boldsymbol{u}_j) r(\boldsymbol{u}_j),$$

where P_i denotes the number of points $\boldsymbol{u}_j$ used to discretize the county v_i, and $e(v_i) = \sum_{j=1}^{P_i} e(\boldsymbol{u}_j)$.

3) 100 realizations of the number of deaths recorded over each county were generated by random drawing of a Poisson distribution whose mean parameter is $r(v_i) \times e(v_i)$.

4) For each realization, we apply the CLMM and PK approaches, using the population-at-risk over the fine grid of 25 km^2 cells as the vector $\boldsymbol{e}_{\mathrm{f}}$ of exposures at the fine resolution.

For all $l = 1, ..., 100$ realizations, the predicted risks $r_{\mathrm{P}}^{(l)}(\boldsymbol{u}_j)$ obtained from both approaches were compared to the underlying risk $r(\boldsymbol{u}_j)$, $j = 1, ..., 3751$, using the following criteria:

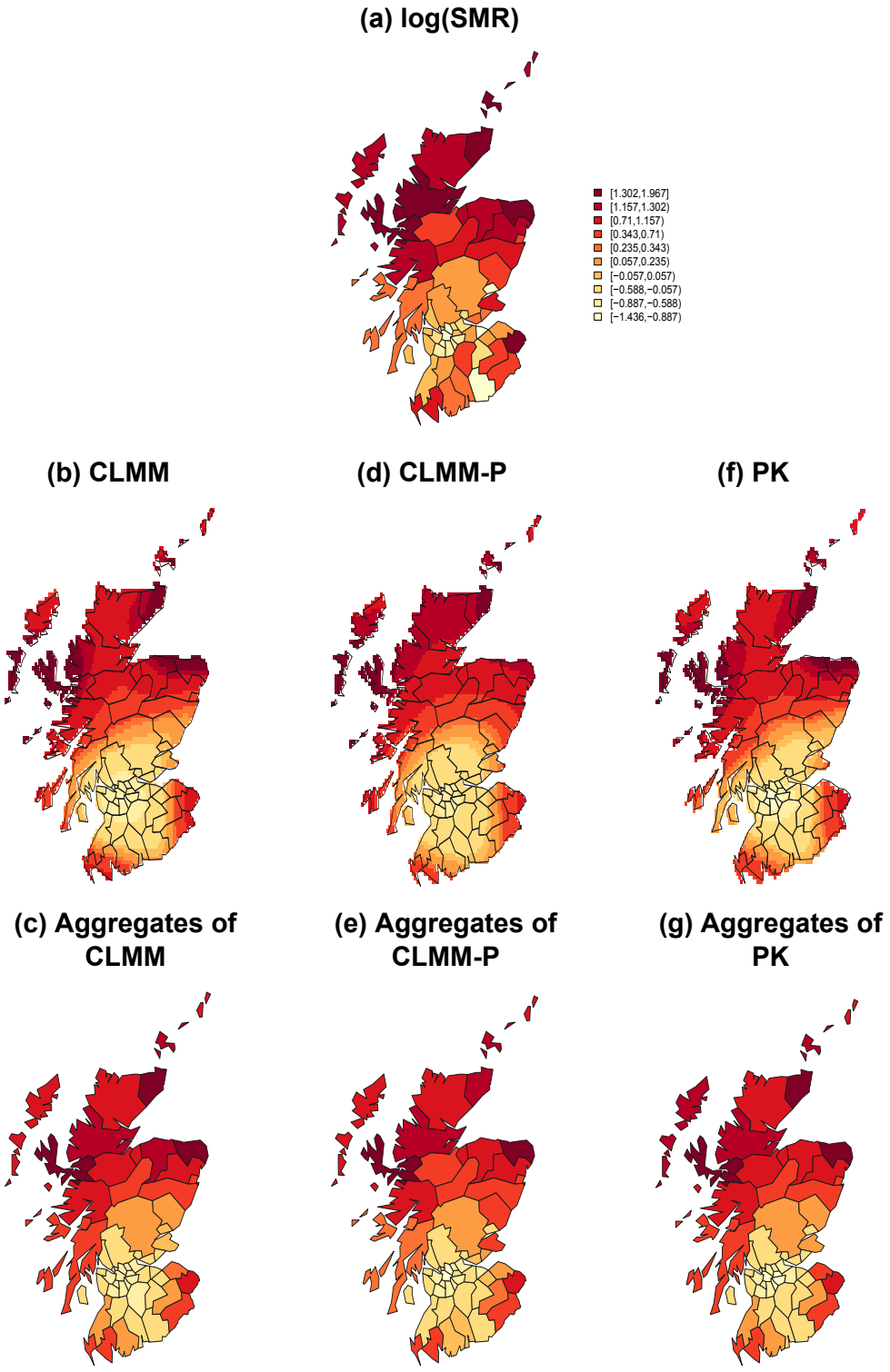


Figure 3.4: Map of (log) standardized mortality rates in Scotland, and the (log) mortality risks estimated by different approaches. The top map shows the log(SMR) recorded over the period 1975-1980 for 56 counties. The middle maps show the smoothed (log) mortality risks at a selected fine grid, which are resulting from the CLMM, CLMM-P, and PK approaches. The bottom maps show the resulting aggregation of these point estimates. The color legend applies to all maps; the class boundaries correspond to the deciles of the log(SMR).

- Mean error (ME):

$$\mathrm{ME}^{(l)} = \frac{1}{W}\sum_{j=1}^{3751} e(\boldsymbol{u}_j)\left(r_{\mathrm{P}}^{(l)}(\boldsymbol{u}_j) - r(\boldsymbol{u}_j)\right) \text{ with } W = \sum_{j=1}^{3751} e(\boldsymbol{u}_j)$$

- Mean absolute error (MAE):

$$\mathrm{MAE}^{(l)} = \frac{1}{W}\sum_{j=1}^{3751} e(\boldsymbol{u}_j)\left|r_{\mathrm{P}}^{(l)}(\boldsymbol{u}_j) - r(\boldsymbol{u}_j)\right| \text{ with } W = \sum_{j=1}^{3751} e(\boldsymbol{u}_j)$$

- Root mean squared error (RMSE):

$$\mathrm{RMSE}^{(l)} = \sqrt{\frac{1}{W}\sum_{j=1}^{3751} e(\boldsymbol{u}_j)\left(r_{\mathrm{P}}^{(l)}(\boldsymbol{u}_j) - r(\boldsymbol{u}_j)\right)^2} \text{ with } W = \sum_{j=1}^{3751} e(\boldsymbol{u}_j)$$

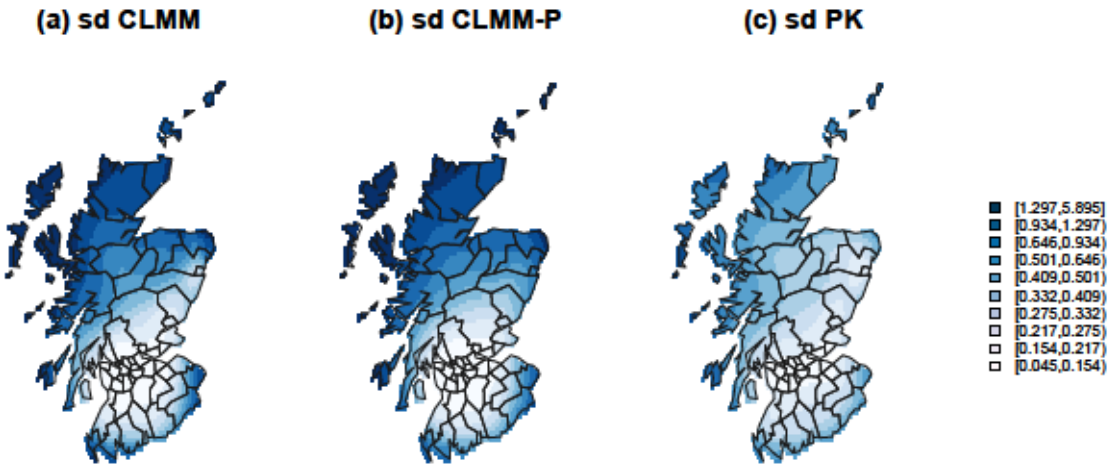


Figure 3.5: Standard error maps for lip cancer incidence in Scotland, estimated by (a) CLMM, (b) CLMM-P and (c) PK approaches.

In all these criteria, the prediction error at each grid point $\boldsymbol{u}_j$ is weighted according to the population size at that location. This was done to penalize more the errors that affect a larger population Goovaerts (2006). Notice that for the ME criterion, it could happen that positive and negative errors are canceled out

so that the true error is underestimated. We have included ME criteria in order to follow the same comparisons as in Goovaerts (2005).

Figure 3.6 shows these resulting errors via box-plots, in which we observe that our approach gives better prediction accuracy than the PK approach, for each criterion. Table 3.1 gives the averages and the standard deviations of the resulting errors (for each criterion) derived from the simulation study. Notice that these results are obtained from a region where the geographical units (the counties) are similar in shape and size. We have conducted an additional simulation study, in which the units vary greatly in shape and size (see Appendix B). For that case, we have considered the Scottish lip cancer dataset. This simulation study shows how the performance of the CLMM is also satisfactory for irregular geographical units.

Approach	ME		MAE		RMSE	
	avg	std	avg	std	avg	std
CLMM	0.0000	0.0006	0.9687	0.0005	1.2553	0.0006
PK	0.0062	0.0005	1.0197	0.0011	1.3514	0.0013

Table 3.1: Performance comparison of CLMM and PK approaches, using different criteria: mean errors (ME), mean absolute errors (MAE), and root mean squared errors (RMSE). These errors are summarized in terms of the average (avg) and standard deviation (std).

3.4 CLMM application to area-to-area case

In the previous section we have seen the application of the spatial CLMM on mortality data recorded at county level, to obtain mortality risk estimates at a desirable fine grid (i.e., the ATP case). Here we illustrate the case when the aim is to obtain such estimates but at a finer geographical unit level than the original one. For that we use a dataset that comes from a large European epidemiological project called MEDEA (see http://www.proyectomedea.org/), whose aim was to study the impact of socio-economic and environmental inequalities on mortality rates by different causes. In particular, we consider deaths by cardiovascular diseases in the Community of Madrid (CM), Spain, recorded at municipality level. Our goal,

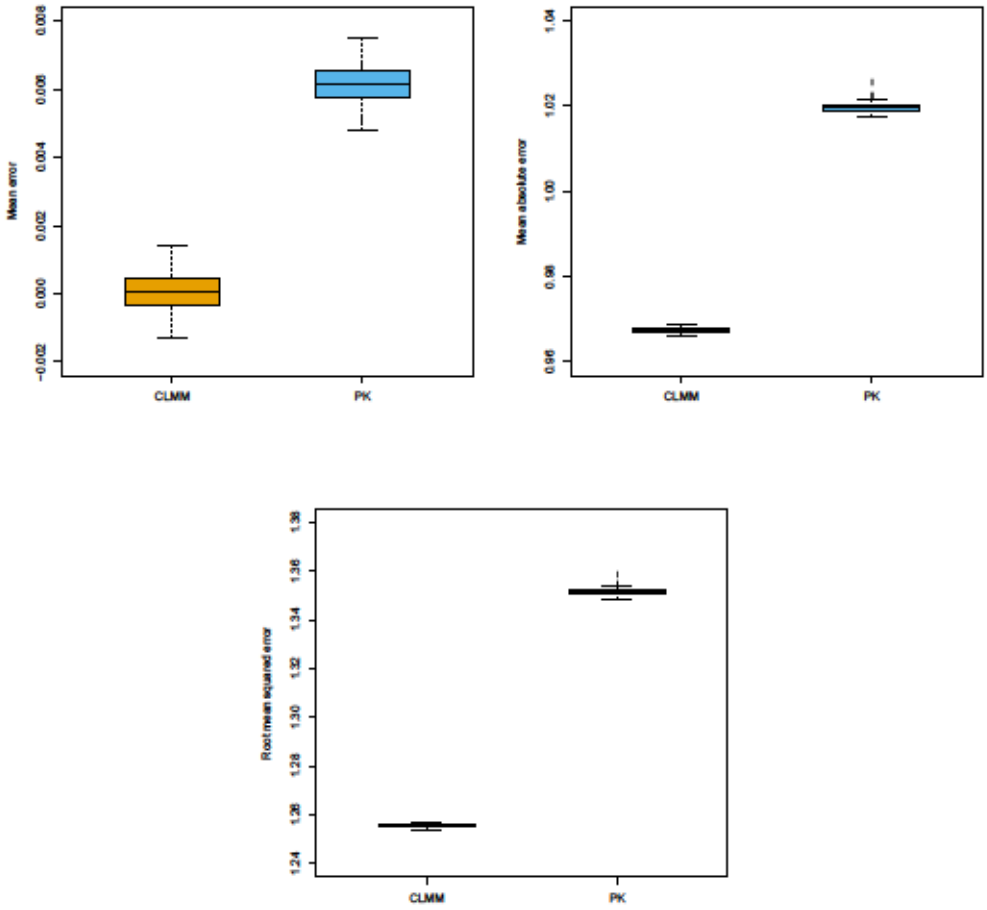


Figure 3.6: Performance comparison between CLMM and PK approaches using different criteria: mean errors (top-left), mean absolute errors (top-right), and root mean squared errors (bottom).

then, is to use the spatial CLMM to obtain mortality risk estimates at census tract level, which is a finer spatial resolution than municipality level.

3.4.1 Raw and smoothed female log(SMRs) at municipality level

Our data correspond to the number of observed and expected female deaths by cardiovascular diseases in the CM over the period 1996-2003. These data were

collected at municipality level, where the geocoding of municipalities corresponds to the year 2001. The cartography used in this section was obtained from the Statistical Institute of the Community of Madrid (see http://www.madrid.org/nomecalles/DescargaBDTCorte.icm).

Figure 3.7 shows the spatial distribution of raw female log(SMR) at municipality level. In 2001, the CM was conformed by 179 municipalities, which vary greatly in shape and size (the largest municipality in terms of surface area is the municipality of Madrid, located at the center of the community). We use a sequential map color scheme with ten equally-weighted classes, where the class boundaries correspond to deciles of raw log(SMR). For 1996-2003 period, the number of observed female deaths varies from 0 to 34884, whereas the number of expected female deaths varies from 0.916 to 44715.610.

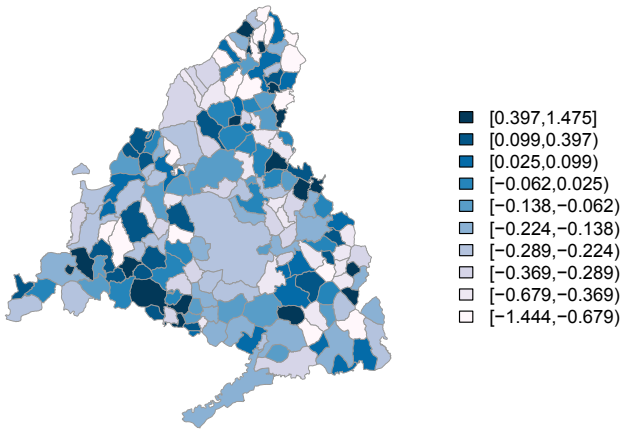


Figure 3.7: Spatial distribution of raw log(SMR) for 179 municipalities of the Community of Madrid over the period 1996-2003. The class boundaries of the color legend correspond to deciles of raw log(SMR).

Since crude log(SMRs) varies abruptly between municipalities, we use the PGLMM approach of Lee and Durbán (2009) to reduce the noise present in these rates and thus to enhance the visualization of underlying trends. For this approach we use the centroid coordinates of the municipalities as covariates, second order penalties, and 20 equally-spaced knots for each marginal cubic B-spline basis. Figure 3.8 shows the resulting smoothed log(SMRs) from PGLMM approach. We observe that most of the higher rates are in the boundaries of the CM, especially in the south-western area. They correspond to areas with difficult access to

health facilities, or industrialized areas where environmental conditions are poor.

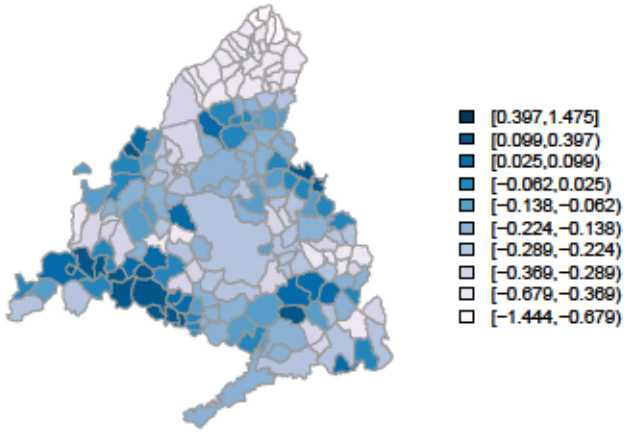


Figure 3.8: Spatial distribution of smoothed log(SMR) for 179 municipalities of the Community of Madrid. The class boundaries of the color legend correspond to deciles of raw log(SMR).

3.4.2 CLMM female log(SMRs) at census tract level

Now, suppose that we seek to visualize the spatial distribution of log(SMR) at census tract level, assuming that we only have mortality data at municipality level. The total number of census tracts for the CM is 3906. To estimate the desired spatial distribution, we use the spatial CLMM for the ATA case, were we must consider the exposures (the number of expected deaths, in this case) at census tract level. For this dataset, we in fact have these quantities, which we denote as $\boldsymbol{e}_{\text{true}}$; otherwise the user has to estimate them in advance. A naive way to do this is to assume that the exposures are evenly distributed throughout the census tracts at each municipality. We denote these resulting estimates as $\widehat{\boldsymbol{e}}_{\text{naive}}$, and we will use them for comparison purpose. The top-left and top-right maps in Figure 3.9 show the resulting smoothed log(SMR) at census tract level, using the ATA CLMM approach (20 equally spaced knots in both directions) with $\boldsymbol{e}_{\text{true}}$ and $\widehat{\boldsymbol{e}}_{\text{naive}}$, respectively. These maps have a similar spatial distribution and are consistent with the smoothed trend obtained at municipality level (see Figure 3.8). The approximate standard errors for these smoothed CLMM log(SMRs) are displayed at the bottom of each mortality map in Figure 3.9. For comparison purpose, we

select the class boundaries for these maps as the cuts of the range of all errors in ten equal parts. We observe that these error maps are very similar and both present high values in the northern area of the CM. The later is due the fact that both smoothed maps are unable to capture more precise mortality trends over the census tracts in this part of the map, where we have less information.

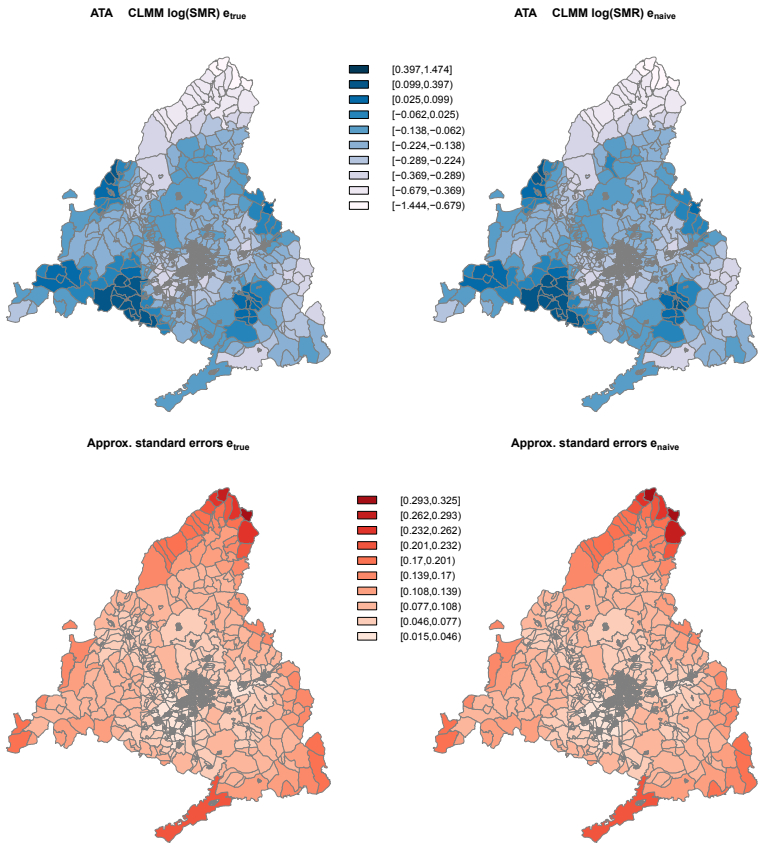


Figure 3.9: Smoothed log(SMR) and their approximate standard errors at census tract level, using the spatial CLMM approach with the true number of expected deaths at census tract level (top-left) and its naive estimator (top-right). The color legend applies to all the maps that show the same quantity; the class boundaries for the smoothed log(SMRs) correspond to the deciles of raw log(SMRs) at municipality level, and the class boundaries for standard errors correspond to the cuts of the range of all errors in ten equal parts.

It is clear that the municipalities of the CM vary greatly in shape and size, especially when we compare the municipality of Madrid (located at the center of the CM) with the rest of them. Figure 3.10 displays the district and census tract boundaries for this municipality, which was conformed by 21 districts and 2358

census tracts in 2001, and the spatial distribution of raw log(SMRs) at district level. The zoom in on the center of this municipality provides a more detailed geographical distribution of the census tracts.

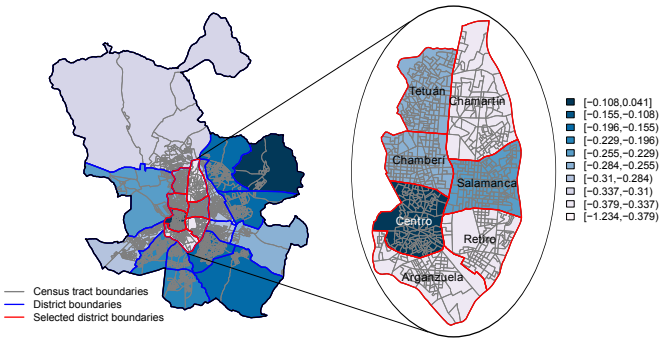


Figure 3.10: Spatial distribution of raw log(SMRs) for the 21 districts in the municipality of Madrid. The zoom shows 7 centric districts of interest and their 780 census tracts. The class boundaries correspond to the deciles of raw log(SMRs) at district level.

Suppose that we only have the number of observed female deaths by cardiovascular diseases for each district in the municipality of Madrid, and we want to estimate mortality rates for the selected districts at census tract level, using the additional information of the number of expected deaths at this level. The CLMM approach for the ATA case is adequate for this situation, in which we want to move from district to census tracts. Figure 3.11 shows the resulting smoothed log(SMRs) using the CLMM approach (20 equally spaced knots for each dimension) with the true vector of exposures. For the area of interest, we observe a more detailed spatial distribution of mortality, where the highest log(SMR) are mostly concentrated around Madrid Centro.

3.4.3 Composite link additive mixed models

In many occasions a researcher may encounter a frequent problem: the covariates to be included in the model are recorded at a different scale of the response variable, or we have different covariates measured at different scales. The CLMM provides a framework in which both situations can be resolved, yielding what we call *composite link additive mixed model*. We start from the premise that in

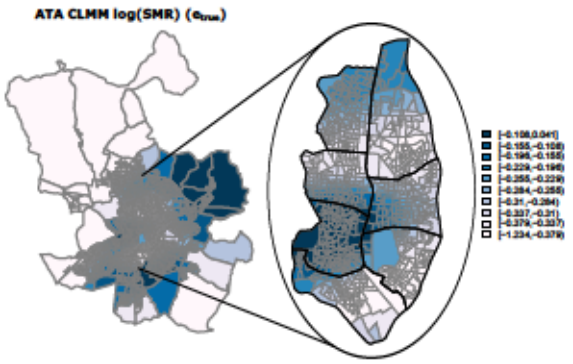


Figure 3.11: Smoothed log(SMRs) using the CLMM approach with the true number of expected deaths at census tract level. The class boundaries for the smoothed log(SMRs) correspond to the deciles of the raw log(SMR) at district level.

all cases we are interested in spatial disaggregation of counts, and so the next situations can occur:

1) If the covariates are measured at fine scale:

$$\boldsymbol{\mu}_{\mathrm{a}} = \mathbf{C}(\boldsymbol{e}_{\mathrm{f}} * \exp(g(\mathrm{space}_{\mathrm{f}})) + \sum_i f_i(\boldsymbol{x}_{\mathrm{if}})),$$

where the subscripts a and f indicate if the variable is measured at the aggregated or fine scale level. g corresponds to a bivariate P-spline introduced in Chapter 3 and f_i could be linear or no linear functions of one or more dimensions of the different covariates. All terms in the model could be represented as mixed models, and therefore the estimation can be carried out as shown previously.

2) If the covariates are measured at the aggregated level:

$$\boldsymbol{\mu}_{\mathrm{a}} = \exp(\log(\mathbf{C}(\boldsymbol{e}_{\mathrm{f}} * \exp(g(\mathrm{space}_{\mathrm{f}})))) + \sum_i f_i(\boldsymbol{x}_{\mathrm{ia}})),$$

this could be seen as an extension of the CLMM-P in which we have a more complicated structure for the random effects.

3) Finally, when covariates are measured at aggregated and fine scale level:

$$\boldsymbol{\mu}_{\mathrm{a}} = \exp(\log(\mathbf{C}(\boldsymbol{e}_{\mathrm{f}} * \exp(g(\mathrm{space}_{\mathrm{f}})) + \sum_j h_j(x_{j\mathrm{f}}))) + \sum_i f_i(\boldsymbol{x}_{\mathrm{ia}}))$$

In practice, the estimation of the functions of the covariates is achieved by extending the mixed model matrices $\mathbf{X}$ and $\mathbf{Z}$ of the vector of latent expectations $\boldsymbol{\gamma}$ (when the covariates are measured at fine scale level). When covariates are measured at aggregated level, the system of equations in Eq. (3.7) would be modified to include the fixed and random effects at the aggregated level.

For illustration, we are going to focus on the first case in which the covariates are measured at the fine scale level. For simplicity, suppose that we want to include a explanatory variable, $\boldsymbol{x}_{\text{l}}$, linearly in the CLMM and another explanatory variable, $\boldsymbol{x}_{\text{nl}}$, in an additive way. Then, the matrices $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{G}^{-1}$ are extended as:

$$\mathbf{X} = [\mathbf{X}_{\text{s}} : \boldsymbol{x}_{\text{l}} : \boldsymbol{x}_{\text{nl}}], \qquad \mathbf{Z} = [\mathbf{Z}_{\text{s}} : \mathbf{Z}_{\text{nl}}], \qquad \mathbf{G}^{-1} = \begin{bmatrix} \mathbf{F}_{\text{s}} & \\ & \lambda_3 \tilde{\boldsymbol{\Sigma}}_3 \end{bmatrix},$$

where the matrices $\mathbf{X}_{\text{s}}$ and $\mathbf{Z}_{\text{s}}$ are considered as in Eq. (3.3) and $\mathbf{F}_{\text{s}}$ as in Eq. (2.50) (which correspond to the spatial part of the model). The matrices $\mathbf{Z}_{\text{nl}}$ and $\tilde{\boldsymbol{\Sigma}}_3$ are computed from $\boldsymbol{x}_{\text{nl}}$ following the mixed model reformulation in the univariate case. The smoothing parameter λ_3 controls the amount of smoothness of the fitted curve for $\boldsymbol{x}_{\text{nl}}$. Then, considering these new matrices, we can use the CLMM estimation procedure in order to obtain optimal estimates for the smoothing parameters λ_1, λ_2, λ_3 (the first two related with the spatial component of the model), and for the fixed and random effects coefficients.

In order to obtain confidence intervals for the fitted curves associated to $\boldsymbol{x}_{\text{l}}$ and $\boldsymbol{x}_{\text{nl}}$, we can use the square root of the variances for each term, which can be computed as (see Ruppert et al., 2003 for more details):

$$\mathbf{K}_j \left(\mathbf{K}'\mathbf{K} + \begin{bmatrix} \mathbf{0} & \\ & \mathbf{G} \end{bmatrix} \right)^{-1} \mathbf{K}'_j,$$

where $\mathbf{K} = [\mathbf{X} : \mathbf{Z}]$, $\mathbf{K}_j = \mathbf{K}\mathbf{E}_j$, and $\mathbf{E}_j$ is a square diagonal matrix that has zeros in the diagonal except in the places that correspond to the j-th term in the CLMM. For example, suppose that the matrix $\mathbf{Z}_{\text{nl}}$ has 13 columns and the matrix $\mathbf{Z}_{\text{s}}$ 221 columns. Also assume that the matrix $\mathbf{X}$ has 6 columns. If we want to compute confidence intervals associated to the fitted curve associated to $\boldsymbol{x}_{\text{nl}}$, the matrix $\mathbf{E}_j$ will have zeros in all the positions except in $\{6, 228, ..., 240\}$.

Now, we illustrate the use of the composite link additive mixed model for the

ATA case. For that, we use two socio-economic explanatory variables related with labour market and employment, which are available at census tract level in the CM, which were also of interest in the MEDEA project in order to stablish the relationship between different socio-economic indicators and mortality rates. The first covariate is an indicator of manual workers whose ages are greater or equal than 16 years old; the second one is an indicator of unemployment for people whose ages are greater or equal than 16 years old. All these indicators are expressed as percentages and were obtained from Census 2001 given by the National Statistics Institute (Instituto Nacional de Estadística, INE). We followed the definitions given in Domínguez-Berjón et al. (2008, p. 182) to construct these indicators. The spatial distribution of percentages of manual workers and unemployed people at census tract level in the CM are depicted in Figure 3.12.

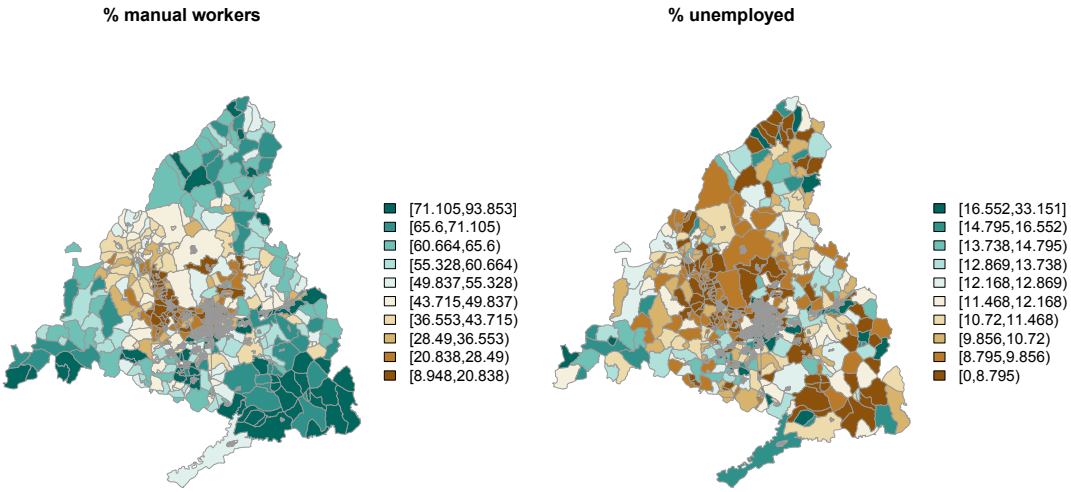


Figure 3.12: Percentage of manual workers (left) and unemployed people (right) greater or equal than 16 years old (the class boundaries correspond to deciles of each percentage).

We use 20 equally spaced knots for the cubic B-spline basis of each spatial coordinate (i.e., the centroids of the census tracts) and 12 equally-spaced knots for the cubic B-spline basis associated with each covariate. The resulting fitted curves for the explanatory variables are shown in Figure 3.13 (effects are centered), where 95% confidence intervals (dashed lines) are also depicted. The relationship between the percentage of manual workers and the mortality rates is linear (the estimated degrees of freedom associated to the estimated curves were approximately one).

As expected, the slope of that line is positive, therefore the higher the level of manual workers, the higher the mortality. This makes sense since the percentage of manual workers is a proxy for low income in the households in the area. On the contrary, the relationship between the percentage of unemployed people and the mortality rates is non-linear. The effect of unemployment is null until we reach 18% percent of unemployment (close to average unemployment rate in Spain). When unemployment is over 20% there is a clear, rapid, and linear increase of mortality rates.

Figure 3.14 shows the remaining spatial effect after accounting for the effects of the covariates. As we can see patterns are similar to those given in Figure 3.9 (when covariates were not included).

Values for AIC for the model with and without covariates were 459.07 and 464.08, respectively, indicating a better goodness of fit when covariates were included. Some work is already been done on adapting the approximate F test used in Wood (2006a) to the case in which covariates are measured at different scale to the one in which the response variable is measured.

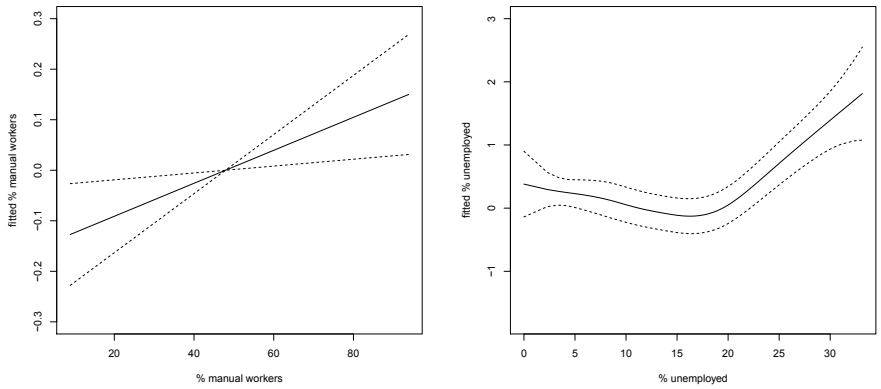


Figure 3.13: Smoothed fitted curves for the covariates: percentage of manual workers and percentage of unemployed people.

3.5 Summary of the chapter

In this chapter we presented the spatial CLMM approach for the area-to-point and area-to-area cases discussed in Section 1.2. For the first case, we obtain

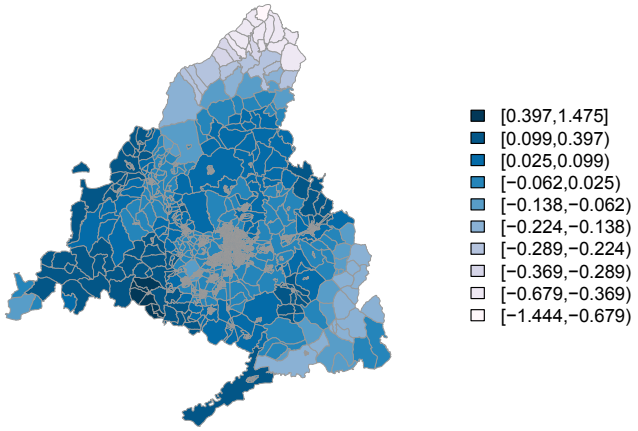


Figure 3.14: Remaining spatial effect after accounting for the effects of the covariates at census tract level.

smoothed estimates that can be depicted in an isopleth map, and for the second one, we obtain smoothed estimates at a finer areal resolution than the observed data resolution. Moreover, we extend the model in order to include covariates that can be measure at different levels of aggregations. We compared our approach with the area-to-point Poisson kriging of Goovaerts (2006) through simulation studies, where geographical units have similar shape and size or not. The simulation results suggest that the spatial CLMM performs better than this geostatistical technique. Moreover, we developed a methodology to deal with the overdispersion problem, which frequently appears when we are working with count data. This methodology, called CLMM-P, was also compared with the previous approaches when the geographical units vary greatly in shape and size.

Chapter 4

Modelling latent spatio-temporal disease incidence with the composite link mixed model approach

This chapter introduces the extension of the CLMM approach to the spatio-temporal case, where count data are grouped in both space and time. In this new context, the mixed model matrices and the composition matrix will contain Kronecker product structures, whose components take into account the spatial and temporal aggregation processes. Also, three smoothing parameters will control the amount of smoothness: two for the spatial coordinates and one for the temporal component. However, this extension leads to the increase in data storage and computational time during the CLMM estimation. To overcome these problems, we propose the use of GLAM algorithms (Currie et al., 2006; Eilers et al., 2006), together with an adaptation of the SAP ('separation of anisotropic penalties') algorithm (Rodríguez-Álvarez et al., 2015) for fast estimation of the amount of smoothness, in the spatio-temporal setting.

This chapter is organized as follows. In Section 4.1 we introduce the CLMM approach for spatio-temporally grouped count data. The use of GLAM methods for the spatio-temporal CLMM is presented in Section 4.3, and the adaptation of the SAP algorithm for our purposes is given in Section 4.2. In Section 4.4 we

apply the CLMM to a Q fever dataset recorded over municipalities (of a specific Dutch study area) and months in 2009, in order to visualize the Q fever incidence on a fine spatial grid along all weeks. In Section 4.5 we conduct a simulation study to examine the performance of the spatio-temporal CLMM approach. Finally, a summary of the chapter is provided in Section 4.6.

4.1 The spatio-temporal composite link mixed model

In order to present the composite link mixed model to the spatio-temporal setting, let first introduce the generalization of the PCLM into this setting.

Let y_{it}, $i = 1, ..., n$, $t = 1, ..., T_1$, denote count data that are recorded over n non-overlapping geographical units, which form the study area of interest, at T_1 time periods. For simplicity, we assume here that the unit boundaries are fixed during these periods (although this assumption can be relaxed under our approach). Now suppose that we want to estimate the latent distribution of the vector of counts $\boldsymbol{y} = (y_{11}, ..., y_{n1}, ..., y_{1T_1}, ..., y_{nT_1})'$ at a spatio-temporal support that is a (nested) refinement of the original one. The fine support is determined by three covariates: $\boldsymbol{x}_1 = (x_{11}, ..., x_{1m})'$ and $\boldsymbol{x}_2 = (x_{21}, ..., x_{2m})'$, with $m > n$, which represent the longitude and latitude coordinates of the spatial refinement, respectively; and $\boldsymbol{x}_3 = (x_{31}, ..., x_{3T_2})'$, with $T_2 > T_1$, which represents the temporal refinement. Assuming that $\boldsymbol{y}$ is distributed Poisson with mean vector $\boldsymbol{\mu} = (\mu_{11}, ..., \mu_{n1}, ..., \mu_{1T_1}, ..., \mu_{nT_1})'$, the spatio-temporal PCLM is given by:

$$\boldsymbol{\mu} = \mathbf{C}_{\text{st}}\boldsymbol{\gamma} = \mathbf{C}_{\text{st}} \exp(\mathbf{B}\boldsymbol{\theta}), \tag{4.1}$$

where $\boldsymbol{\gamma}$ denotes the latent mean at the fine resolution, $\mathbf{C}_{\text{st}}$ is the new spatio-temporal composition matrix, and $\mathbf{B}$ is a full regression basis constructed from covariates $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_3$. To achieve smoothness, the vector of regression coefficients $\boldsymbol{\theta}$ is penalized by a discrete penalty matrix $\mathbf{P}$ in the form $\boldsymbol{\theta}'\mathbf{P}\boldsymbol{\theta}$. Matrices $\mathbf{B}$, $\mathbf{P}$, and $\mathbf{C}_{\text{st}}$ in Eq. (4.1) are described below.

Following the proposal of Lee and Durbán (2011), the regression basis $\mathbf{B}$ in

Eq. (4.1) is given by:

$$\mathbf{B} = \mathbf{B}_3 \otimes (\mathbf{B}_2 \square \mathbf{B}_1)\,, \tag{4.2}$$

where $\mathbf{B}_1 = \mathbf{B}(\boldsymbol{x}_1)$, $\mathbf{B}_2 = \mathbf{B}(\boldsymbol{x}_2)$, and $\mathbf{B}_3 = \mathbf{B}(\boldsymbol{x}_3)$ are marginal B-spline bases of dimensions $m \times c_1$, $m \times c_2$ and $T_2 \times c_3$, respectively. For the spatio-temporal case, Lee and Durbán (2011) propose to use the three-dimensional discrete penalty matrix given in Eq. (2.47). Notice that this penalty matrix allows for anisotropy, i.e., a different amount of smoothing for $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$, which is a desirable characteristic in the spatio-temporal context.

The spatio-temporal composition matrix $\mathbf{C}_{\text{st}}$ in Eq. (4.1) can be expressed as:

$$\mathbf{C}_{\text{st}} = \mathbf{C}_{\text{t}} \otimes \mathbf{C}_{\text{s}}, \tag{4.3}$$

where $\mathbf{C}_{\text{s}}$ and $\mathbf{C}_{\text{t}}$ are the spatial and temporal composition matrices of dimensions $n \times m$ and $T_1 \times T_2$, respectively. The definition of these matrices depends on the type and level of aggregation. For example, if we want to estimate the latent distribution at a fine spatial grid (over the study area), the entries of the spatial composition matrix $\mathbf{C}_{\text{s}}$ can be computed as in Eq. (3.5). The temporal composition matrix $\mathbf{C}_{\text{t}}$ can be used to disaggregate coarse time intervals into detailed time periods (for example, from years or trimesters to months, weeks, or days). In Section 4.2, we show the structure that $\mathbf{C}_{\text{t}}$ will have, specifically for our purposes. Notice that if $\mathbf{C}_{\text{s}} = \mathbf{I}_n$, $\mathbf{C}_{\text{t}} = \mathbf{I}_{T_1}$, and the unit centroids are used as spatial covariates, the presented methodology correspond to a Poisson version of the proposal given by Lee and Durbán (2011) (i.e, the PGLMM approach), for the smoothing of spatio-temporal count data.

When spatial data are recorded over a coarse regular grid, a more appropriate definition for the regression basis $\mathbf{B}$ in Eq. (4.1) is $\mathbf{B} = \mathbf{B}_3 \otimes (\mathbf{B}_2 \otimes \mathbf{B}_1)$, where the spatial refinement correspond to the cell centroid coordinates of a fine grid. Moreover, the spatial composition matrix can be written as $\mathbf{C}_{\text{s}} = \mathbf{C}_2 \otimes \mathbf{C}_1$, where each $\mathbf{C}_d$, $d = 1, 2$, is constructed according to the disaggregation of the coarser grid cells into small ones. In this sense, the spatio-temporal PCLM will be identical to the three-dimensional PCLM for coarse array data given in Section 2.3.

The spatio-temporal composite link mixed model is obtained by reformulating the model in Eq. (4.1), subject to the penalization given in Eq. (2.47), as a mixed

model:

$$\boldsymbol{\mu} = \mathbf{C}_{\text{st}}\boldsymbol{\gamma} = \mathbf{C}_{\text{st}}\left(\boldsymbol{e}_{\text{f}} * \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha})\right), \text{ with } \boldsymbol{\alpha} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{G}\right), \tag{4.4}$$

where population information at the fine scale is incorporated by means of the vector $\boldsymbol{e}_{\text{f}}$. Following the proposal of Lee and Durbán (2011), the mixed model matrices $\mathbf{X}$ and $\mathbf{Z}$ of the model in Eq. (4.4) are obtained as:

$$\begin{aligned}
\mathbf{X} &= \mathbf{X}_3 \otimes (\mathbf{X}_2 \square \mathbf{X}_1), \\
\mathbf{Z} &= [\mathbf{Z}_3 \otimes (\mathbf{X}_2 \square \mathbf{X}_1) : \mathbf{X}_3 \otimes (\mathbf{Z}_2 \square \mathbf{X}_1) : \mathbf{X}_3 \otimes (\mathbf{X}_2 \square \mathbf{Z}_1) : \\
&\quad \mathbf{Z}_3 \otimes (\mathbf{Z}_2 \square \mathbf{X}_1) : \mathbf{Z}_3 \otimes (\mathbf{X}_2 \square \mathbf{Z}_1) : \mathbf{X}_3 \otimes (\mathbf{Z}_2 \square \mathbf{Z}_1) : \mathbf{Z}_3 \otimes (\mathbf{Z}_2 \square \mathbf{Z}_1)],
\end{aligned} \tag{4.5}$$

where $\mathbf{X}_d = \mathbf{B}_d\mathbf{U}_{dn}$ and $\mathbf{Z}_d = \mathbf{B}_d\mathbf{U}_{ds}$, for $d = 1, 2, 3$, with $\mathbf{U}_{dn}$ and $\mathbf{U}_{ds}$ defined as in Section 2.3. Moreover, the inverse of the covariance matrix $\mathbf{G}$ of the random effects $\boldsymbol{\alpha}$ in Eq. (4.4) becomes the block-diagonal matrix $\mathbf{F}$ given in (2.53), with smoothing parameters λ_1 and λ_2 for the spatial coordinates, and a smoothing parameter λ_3 for the temporal component.

Regarding the parameter estimation of the spatio-temporal CLMM in Eq. (4.4), it can be carried out using the procedure given in Section 2.2. That is, for fixed values of λ_1, λ_2, and λ_3, the estimation of the fixed and random effects $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ in Eq. (4.4) are obtained by using the PQL approach (see Eq. (2.33)), whereas the smoothing parameters can be estimated by numerically maximizing the approximate REML given in (2.37). In this case, however, the use of numerical optimization methods for REML criterion is not efficient, leading to the increase of the computational burden. Moreover, the direct computation of matrix cross-products in which the mixed model matrices $\mathbf{X}$ and $\mathbf{Z}$ (defined in Eq. (4.5)) and the spatio-temporal composition matrix (defined in Eq. (4.3)) are involved, can easily lead to storage problems.

In the next sections we provide solutions for the CLMM estimation in terms of storage and efficiency. In Section 4.2 we adapt the SAP algorithm, given by Rodríguez-Álvarez et al. (2015), to the spatio-temporal CLMM context in order to efficiently compute estimates for the smoothing parameters. To avoid possible storage problems, in Section 4.3 we adapt the GLAM algorithms (Currie et al., 2006; Eilers et al., 2006) in the spatio-temporal CLMM setting. These algorithms

also offer an efficient way to compute required matrix cross-products in the SAP algorithm (such as, for example, $\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}}$ and $\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{X}}$), alleviating the computing time of the model estimation.

4.2 SAP algorithm for spatio-temporal CLMMs

Under the CLMM approach (see Section 2.2), and conditioning on the estimators given in Eq. (2.33), we can obtain estimates for λ_1, λ_2, and λ_3 by numerically maximizing the approximate REML given in (2.37). To avoid the use of any numerical optimization method, in this section we adapt the SAP algorithm of Rodríguez-Álvarez et al. (2015) to the CLMM setting. This algorithm is a generalization of the work by Schall (1991) that deals with non-standard structures of the covariance matrix of the random effects, where the smoothing parameters are seen as ratios of variance components, i.e., $\lambda_d = \frac{\phi}{\tau_d^2}$, for $d = 1, 2, 3$. Since we are working under a Poisson framework, $\phi = 1$. Thus, the problem is reduced to obtain estimates for the variance components τ_1^2, τ_2^2, and τ_3^2.

Following the proposal of Rodríguez-Álvarez et al. (2015), we can derive closed-form expressions for the REML estimates of the variance components τ_d^2 (for $d = 1, 2, 3$). These estimates are given by:

$$\widehat{\tau}_d^2 = \frac{\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Lambda}_d\widehat{\boldsymbol{\alpha}}}{\mathrm{ed}_d}, \tag{4.6}$$

where:

$$\mathrm{ed}_d = \mathrm{trace}\left(\breve{\mathbf{Z}}'\mathbf{N}\breve{\mathbf{Z}}\mathbf{G}\frac{\boldsymbol{\Lambda}_d}{\tau_d^2}\mathbf{G}\right), \tag{4.7}$$

with $\mathbf{N}$ defined as in Eq. (2.36), and

$$\begin{aligned}
\boldsymbol{\Lambda}_1 &= \mathrm{blockdiag}\left(\mathbf{0}_{q_1q_2(c_3-q_3)}, \mathbf{0}_{q_1q_3(c_2-q_2)}, \mathbf{F}_{1u}, \mathbf{0}_{q_1(c_2-q_2)(c_3-q_3)}, \mathbf{F}_{12}, \mathbf{F}_{11}, \mathbf{F}_{1t}\right), \\
\boldsymbol{\Lambda}_2 &= \mathrm{blockdiag}\left(\mathbf{0}_{q_1q_2(c_3-q_3)}, \mathbf{F}_{2u}, \mathbf{0}_{q_2q_3(c_1-q_1)}, \mathbf{F}_{22}, \mathbf{0}_{q_2(c_1-q_1)(c_3-q_3)}, \mathbf{F}_{21}, \mathbf{F}_{2t}\right), \\
\boldsymbol{\Lambda}_3 &= \mathrm{blockdiag}\left(\mathbf{F}_{3u}, \mathbf{0}_{q_1q_3(c_2-q_2)}, \mathbf{0}_{q_2q_3(c_1-q_1)}, \mathbf{F}_{32}, \mathbf{F}_{31}, \mathbf{0}_{q_3(c_1-q_1)(c_2-q_2)}, \mathbf{F}_{3t}\right).
\end{aligned}$$

The non-null submatrices of each Λ_d, $d = 1, 2, 3$, were previously defined in Section 2.3. The proof of this result is provided in Appendix C. Notice that the inverse of the covariance matrix $\mathbf{G}$ in Eq. (4.4) can be decomposed as $\mathbf{G}^{-1} =$

$\frac{1}{\tau_1^2}\mathbf{\Lambda}_1 + \frac{1}{\tau_2^2}\mathbf{\Lambda}_2 + \frac{1}{\tau_3^2}\mathbf{\Lambda}_3$, where the capital lambdas are defined above. An algorithm for the CLMM estimation (which is an adaptation of the algorithm provided in Rodríguez-Álvarez et al., 2015, p. 945) is given in Algorithm 1.

The computation of the trace given in Eq. (4.7) can be efficiently obtained by taking into account that $\mathbf{G}\mathbf{\Lambda}_d\mathbf{G}$ is a diagonal matrix. Thus, we only need to compute the diagonal of $\breve{\mathbf{Z}}'\mathbf{N}\breve{\mathbf{Z}}$ in order to obtain this trace. From Harville (1977, Eq. (5.3)) we have that:

$$\breve{\mathbf{Z}}'\mathbf{N} = [\mathbf{0}_{(c_1c_2c_3-q_1q_2q_3)\times q_1q_2q_3}|\mathbf{I}_{(c_1c_2c_3-q_1q_2q_3)}]\mathbf{M}_*^{-1}[\breve{\mathbf{X}}|\breve{\mathbf{Z}}]\mathbf{W},$$

where:

$$\mathbf{M}_* = \begin{bmatrix} \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{X}} & \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}}\mathbf{G} \\ \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{X}} & \mathbf{I} + \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}}\mathbf{G} \end{bmatrix}.$$

Therefore, the diagonal elements of the matrix $\breve{\mathbf{Z}}'\mathbf{N}\breve{\mathbf{Z}}$ are obtained by the column-wise addition of

$$\left(\mathbf{0}_{(c_1c_2c_3-q_1q_2q_3)\times q_1q_2q_3}|\mathbf{I}_{(c_1c_2c_3-q_1q_2q_3)}]\mathbf{M}_*^{-1}\right)' \odot \begin{bmatrix} \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}} \\ \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}} \end{bmatrix}.$$

Notice that the matrix cross-products $\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}}$, $\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}}$, and $\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{X}}$ are involved in the previous column-wise addition. These matrix cross-products also appear in the estimation of the fixed and random coefficients, which can be obtained by solving the linear system given in Eq. (2.32). We can efficiently compute them by adapting GLAM algorithms in the CLMM setting, which we will present in the next section.

Using the SAP algorithm for the estimation of the model in Eq. (4.4), we can directly compute the effective dimension of this model as:

$$\text{ED} = \sum_{d=1}^{3} \text{ed}_d + q_1q_2q_3, \tag{4.8}$$

where the ed_d expressions are computed from Eq. (4.7). The first term in the right-hand side of Eq. (4.8) corresponds to the dimension of the penalized part (see Appendix C), whereas the second to the unpenalized part.

Algorithm 1 SAP algorithm for CLMM's parameters estimation

Require: Convergence tolerances ν_1 and ν_2 (e.g., 1×10^{-6}), and maximum number of iterations maxit_1 and maxit_2 (e.g., 100).

1: Set initial values for the mixed model coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, and the variance components τ_1^2, τ_2^2, and τ_3^2 (for example, $\widehat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$ with length $q_1q_2q_3$, $\widehat{\boldsymbol{\alpha}}^{(0)} = \mathbf{0}$ with length $(c_1c_2c_3 - q_1q_2q_3)$, and $\widehat{\tau}_1^{2(0)} = \widehat{\tau}_2^{2(0)} = \widehat{\tau}_3^{2(0)} = 1$). Set $k = 0$

2: **for** 1 **to** maxit_1 **do**

3: Given the current mixed model coefficients' estimates, construct the matrix of weights $\mathbf{W}$ and the working vector $\boldsymbol{z}$ as follows:

$$\boldsymbol{\Gamma} = \text{diag}(\widehat{\boldsymbol{\gamma}}^{(k)}) \text{ , with } \widehat{\boldsymbol{\gamma}}^{(k)} = \boldsymbol{e}_{\text{f}} * \exp(\mathbf{X}\widehat{\boldsymbol{\beta}}^{(k)} + \mathbf{Z}\widehat{\boldsymbol{\alpha}}^{(k)});$$
$$\mathbf{W} = \text{diag}(\widehat{\boldsymbol{\mu}}^{(k)}) \text{ , with } \widehat{\boldsymbol{\mu}}^{(k)} = \mathbf{C}\widehat{\boldsymbol{\gamma}}^{(k)};$$
$$\boldsymbol{z} = \breve{\mathbf{X}}\widehat{\boldsymbol{\beta}}^{(k)} + \breve{\mathbf{Z}}\widehat{\boldsymbol{\alpha}}^{(k)} + \mathbf{W}^{-1}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}^{(k)}) = \mathbf{W}^{-1}\mathbf{C}\boldsymbol{\Gamma}\widehat{\boldsymbol{\eta}}^{(k)} + \mathbf{W}^{-1}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}^{(k)}),$$

with $\widehat{\boldsymbol{\eta}}^{(k)} = \mathbf{X}\widehat{\boldsymbol{\beta}}^{(k)} + \mathbf{Z}\widehat{\boldsymbol{\alpha}}^{(k)}$.

4: **for** 1 **to** maxit_2 **do**

5: Given the current estimates for the variance components, obtain new estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ by solving the linear system given in Eq. (2.32). The resulting estimates are denoted as $\widehat{\boldsymbol{\beta}}^{(k+1)}$ and $\widehat{\boldsymbol{\alpha}}^{(k+1)}$, respectively.

6: Obtain new estimates for the variance components from Eq. (4.6). The resulting estimates are denoted as $\widehat{\tau}_d^{2(k+1)}$, for $d = 1, 2, 3$.

7: Compare new variance components' estimates with the previous ones, using the following convergence criterion:

$$\frac{\sum_{d=1}^{3} |\widehat{\tau}_d^{2(k+1)} - \widehat{\tau}_d^{2(k)}|}{3} \leq \nu_1.$$

8: If the convergence tolerance is achieved, **break**, otherwise set $\widehat{\tau}_d^{2(k)} = \widehat{\tau}_d^{2(k+1)}$ and repeat steps 5, 6, and 7 until convergence.

9: **end for**

10: Compute new estimates for the vector of smooth trends at fine scale using the model's fixed and random effects estimates obtained in the last iteration of step 5. The resulting vector is denoted as $\widehat{\boldsymbol{\eta}}^{(k+1)}$. Compare new estimates with the previous ones, using the following convergence criterion:

$$\frac{\|\widehat{\boldsymbol{\eta}}^{(k+1)} - \widehat{\boldsymbol{\eta}}^{(k)}\|^2}{\|\widehat{\boldsymbol{\eta}}^{(k+1)}\|^2} \leq \nu_2,$$

11: If the convergence tolerance is achieved, **break**, otherwise set $k = k+1$ and repeat the previous steps until convergence.

12: **end for**

4.3 GLAM methods for spatio-temporal CLMMs

When we are dealing with the estimation of latent trends in multiple dimensions, we are susceptible to encounter problems with storage and computational burden. As we seen in Section 2.3 for the case of data arranged in multidimensional grids, it is possible to circumvent these problems using GLAM methods developed in Currie et al. (2006) and Eilers et al. (2006). These methods are designed to avoid the direct computation of matrix cross-products where Kronecker operations are involved, by using sequences of nested matrix operations. In this section we show the use of these methods in the spatio-temporal CLMM context.

In the previous section we have seen that several matrix cross-products have to be computed as, for example, $\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}}$ and $\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}}$. Also, another matrix cross-products, such as $\breve{\mathbf{X}}'\mathbf{W}\boldsymbol{z}$ and $\breve{\mathbf{Z}}'\mathbf{W}\boldsymbol{z}$, are needed to obtain estimates for the fixed and the random effects coefficients (see Eq. (2.32) and line 5 in Algorithm 1). The GLAM methods offer a fast and an efficient way to compute them as follows.

First, the matrix-by-vector products that we need to compute are: $\mathbf{X}\boldsymbol{\beta}$, $\mathbf{Z}\boldsymbol{\alpha}$, and $\mathbf{C}\boldsymbol{\gamma}$ (see line 3 in Algorithm 1). These expressions are computed as:

$$
\begin{aligned}
&\mathbf{X}\boldsymbol{\beta} \equiv \rho(\mathbf{X}_3, \rho(\mathbf{R}_1, \tilde{\mathbf{Q}})),\\
&\mathbf{Z}\boldsymbol{\alpha} \equiv \rho(\mathbf{Z}_3, \rho(\mathbf{R}_1, \tilde{\mathbf{A}}_1)) + \rho(\mathbf{X}_3, \rho(\mathbf{R}_2, \tilde{\mathbf{A}}_2)) + \rho(\mathbf{X}_3, \rho(\mathbf{R}_3, \tilde{\mathbf{A}}_3)) + \rho(\mathbf{Z}_3, \rho(\mathbf{R}_2, \tilde{\mathbf{A}}_4)) +\\
&\qquad \rho(\mathbf{Z}_3, \rho(\mathbf{R}_3, \tilde{\mathbf{A}}_5)) + \rho(\mathbf{X}_3, \rho(\mathbf{R}_4, \tilde{\mathbf{A}}_6)) + \rho(\mathbf{Z}_3, \rho(\mathbf{R}_4, \tilde{\mathbf{A}}_7)),\\
&\mathbf{C}\boldsymbol{\gamma} \equiv \rho(\mathbf{C}_t, \rho(\mathbf{C}_s, \tilde{\boldsymbol{\Gamma}})),
\end{aligned}
$$

with $\mathbf{R}_1 = \mathcal{G}(\mathbf{X}_2, \mathbf{X}_1)$, $\mathbf{R}_2 = \mathcal{G}(\mathbf{Z}_2, \mathbf{X}_1)$, $\mathbf{R}_3 = \mathcal{G}(\mathbf{X}_2, \mathbf{Z}_1)$, $\mathbf{R}_4 = \mathcal{G}(\mathbf{Z}_2, \mathbf{Z}_1)$, where ρ and $\mathcal{G}$ denote the rotated $\mathcal{H}-$transform and the row-tensor product, respectively, which are defined in Appendix A, and the symbol $\equiv$ means that both sides have the same elements but in a different order. The matrices $\tilde{\mathbf{Q}}$, $\tilde{\boldsymbol{\Gamma}}$, and the $\tilde{\mathbf{A}}_k$'s (for $k = 1, ..., 7$) are arrangements of the vectors $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\alpha}_k$'s, respectively, with $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1', ..., \boldsymbol{\alpha}_7')'$, whose dimensions correspond to the number of columns of the first matrix where ρ acts, times the number of columns of the second matrix where ρ acts (i.e, $\tilde{\mathbf{Q}}$ has dimension $\text{ncol}(\mathbf{R}_1) \times \text{ncol}(\mathbf{X}_3) = q_1 q_2 \times q_3$, $\tilde{\boldsymbol{\Gamma}}$ has dimension $\text{ncol}(\mathbf{C}_s) \times \text{ncol}(\mathbf{C}_t) = m \times T_2$, and so on). Therefore, it holds that $\text{vec}(\tilde{\mathbf{Q}}) = \boldsymbol{\beta}$, $\text{vec}(\tilde{\boldsymbol{\Gamma}}) = \boldsymbol{\gamma}$, and $\text{vec}(\tilde{\mathbf{A}}_k) = \boldsymbol{\alpha}_k$, for $k = 1, ..., 7$.

Also, the following matrix cross-products are needed: $\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{X}}$, $\breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}}$, $\breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}}$,

$\check{\mathbf{Z}}'\mathbf{W}\check{\mathbf{X}}$ (which is equal to $(\check{\mathbf{X}}'\mathbf{W}\check{\mathbf{Z}})'$), $\check{\mathbf{X}}'\mathbf{W}\boldsymbol{z}$, and $\check{\mathbf{Z}}'\mathbf{W}\boldsymbol{z}$. First, note that they can be reduced as:

$$\begin{aligned}
&\check{\mathbf{X}}'\mathbf{W}\check{\mathbf{X}} = (\mathbf{C}\boldsymbol{\Gamma}\mathbf{X})'\mathbf{W}^{-1}(\mathbf{C}\boldsymbol{\Gamma}\mathbf{X}), \quad \check{\mathbf{Z}}'\mathbf{W}\check{\mathbf{Z}} = (\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z})'\mathbf{W}^{-1}(\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z}),\\
&\check{\mathbf{X}}'\mathbf{W}\check{\mathbf{Z}} = (\mathbf{C}\boldsymbol{\Gamma}\mathbf{X})'\mathbf{W}^{-1}(\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z}), \quad \check{\mathbf{X}}'\mathbf{W}\boldsymbol{z} = (\mathbf{C}\boldsymbol{\Gamma}\mathbf{X})'\boldsymbol{z},\\
&\check{\mathbf{Z}}'\mathbf{W}\boldsymbol{z} = (\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z})'\boldsymbol{z}.
\end{aligned}$$

Thus, we only need to compute $\mathbf{C}\boldsymbol{\Gamma}\mathbf{X}$ and $\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z}$. These expressions are computed as follows:

$$\begin{aligned}
\mathbf{C}\boldsymbol{\Gamma}\mathbf{X} &\equiv \rho(\mathcal{G}(\mathbf{X}_3, \mathbf{C}_t')', \rho(\mathcal{G}(\mathbf{R}_1, \mathbf{C}_s')', \tilde{\boldsymbol{\Gamma}})),\\
\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z} &\equiv [\rho(\mathcal{G}(\mathbf{Z}_3, \mathbf{C}_t')', \rho(\mathcal{G}(\mathbf{R}_1, \mathbf{C}_s')', \tilde{\boldsymbol{\Gamma}})) : \rho(\mathcal{G}(\mathbf{X}_3, \mathbf{C}_t')', \rho(\mathcal{G}(\mathbf{R}_2, \mathbf{C}_s')', \tilde{\boldsymbol{\Gamma}})) :\\
&\quad \rho(\mathcal{G}(\mathbf{X}_3, \mathbf{C}_t')', \rho(\mathcal{G}(\mathbf{R}_3, \mathbf{C}_s')', \tilde{\boldsymbol{\Gamma}})) : \rho(\mathcal{G}(\mathbf{Z}_3, \mathbf{C}_t')', \rho(\mathcal{G}(\mathbf{R}_2, \mathbf{C}_s')', \tilde{\boldsymbol{\Gamma}})) :\\
&\quad \rho(\mathcal{G}(\mathbf{Z}_3, \mathbf{C}_t')', \rho(\mathcal{G}(\mathbf{R}_3, \mathbf{C}_s')', \tilde{\boldsymbol{\Gamma}})) : \rho(\mathcal{G}(\mathbf{X}_3, \mathbf{C}_t')', \rho(\mathcal{G}(\mathbf{R}_4, \mathbf{C}_s')', \tilde{\boldsymbol{\Gamma}})) :\\
&\quad \rho(\mathcal{G}(\mathbf{Z}_3, \mathbf{C}_t')', \rho(\mathcal{G}(\mathbf{R}_4, \mathbf{C}_s')', \tilde{\boldsymbol{\Gamma}}))]
\end{aligned}$$

with $\tilde{\boldsymbol{\Gamma}}$ defined above.

As we have seen above, the GLAM methods use several rearrangement and redimensioning operations, but these are efficient operations in comparison to the Kronecker product. Table 4.1 gives some comparatives on timing for the computation of the matrix products $\mathbf{C}\boldsymbol{\Gamma}\mathbf{X}$ and $\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z}$, with and without using GLAM methods. These computations were performed on a 1.80 GHz Intel® Core™ i7 processor computer with 4 GB of RAM and Windows® 8.1 operating system, using hypothetical matrices $\mathbf{C}$, $\boldsymbol{\Gamma}$, $\mathbf{X}$, and $\mathbf{Z}$, of dimension 40×165160, 165160×165160, 165160×8, and 165160×839, respectively. In both cases the use of GLAM methods improve the computational burden.

Matrix product	Times (s) without GLAM	Times (s) with GLAM	Ratio
$\mathbf{C}\boldsymbol{\Gamma}\mathbf{X}$	1.95	0.08	24 : 1
$\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z}$	183.09	8.74	21 : 1

Table 4.1: User CPU times to calculate $\mathbf{C}\boldsymbol{\Gamma}\mathbf{X}$ and $\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z}$.

4.4 Application: Q fever outbreak in the Netherlands

In this section we illustrate our proposal using data related with Q fever outbreaks in the Netherlands. First we briefly describe these data, and then we analyse them using the spatio-temporal CLMM approach described in Section 4.1.

4.4.1 Q fever data

Q fever is a widespread zoonotic disease caused by the bacterium *Coxiella burnetii*. The transmission to humans of *C. burnetii* is primary associated with ruminants like cattle, sheep, and goats. During parturition or abortion of infected animals, high numbers of *C. burnetii* are shed within the amniotic fluids and the placenta. These organisms end up in the environment where they may survive for long periods of time due by their resistant to heat, drying, and many common disinfectants. Humans are often very susceptible to the disease, and very few organisms may be required to cause infection. More information about this infectious disease is provided in Maurin and Raoult (1999).

The south of the Netherlands faced large outbreaks of human Q fever from 2007 to 2010. In this country, local municipal health services (MHSs) are responsible for recording every confirmed diagnosis of acute Q fever. The information collected is then entered into the electronic national infectious diseases surveillance database. Due to confidentiality, these data are not publicly available and, in some cases, may be provided in an aggregated form.

Figure 4.1 shows the temporal distribution of Q fever cases (in months) from January 2007 to July 2010. A total of 3807 acute Q fever cases were registered in this period: 192 in 2007, 980 in 2008, 2309 in 2009, and 325 in 2010. The epidemic peaks of each year are observed every spring, specifically during May. This coincides with small ruminants (sheep and goats) birth period — a fact that was pointed out in several studies about this exceptionally large Q fever outbreaks in the Netherlands (see, for example, van der Hoek et al., 2010; Roest et al., 2011). Since the largest outbreak was observed during 2009, we will study the distribution of Q fever incidence in this year.

Figure 4.2a shows the geographical distribution of the residential addresses of

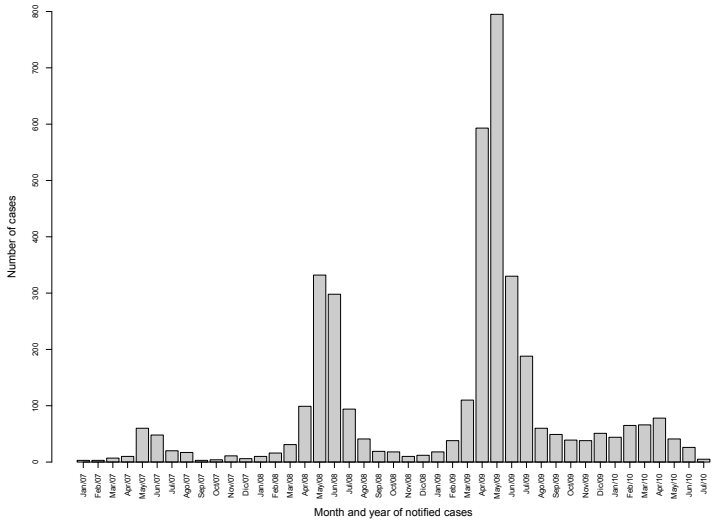


Figure 4.1: Human Q fever cases in the Netherlands grouped per months, from January 2007 to July 2010.

human Q fever cases (red points) in 2009. If we select a 60 × 60 km area in the south of the Netherlands (see black square in Figure 4.2a), 72 municipalities overlap with this study area. The total number of Q fever cases reported in these municipalities is 1798. Aggregating these cases per municipality, and taking into account the number of inhabitants of each municipality, we can calculate the Q fever incidence (per 100000 inhabitants). Figure 4.2b shows the spatial distribution of the resulting Q fever incidence, where higher incidence values are observed around the municipalities of Landerd (1439.676), Lith (562.546), and Heusden (295.006).

4.4.2 Detailed smooth incidence maps

Figure 4.2b shows the choropleth map of raw Q fever incidences at municipality level. However, our aim is to estimate latent incidence maps at detailed periods of time, in order to obtain a better insight of the evolution of the incidence. The spatio-temporal CLMM approach, developed in Section 4.1, allows to visualize the Q fever incidence at a finer spatio-temporal resolution, and also to incorporate population information at fine scale into the estimation of the latent process. Here, we illustrate the application of our methodology using Q fever data collected

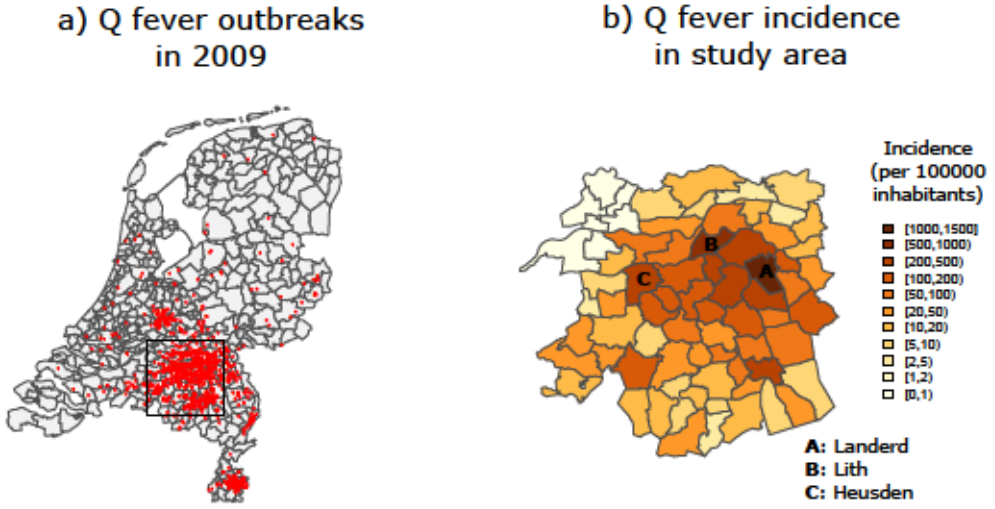


Figure 4.2: Map of human Q fever cases in the Netherlands, 2009. Left: the red points indicate the residential addresses of human cases (2309 in total). Right: study area in the south of the Netherlands showing (crude) incidence (per 100000 inhabitants) of Q fever by municipality in 2009.

over municipalities and months in 2009. Our goal is to obtain Q fever incidence estimates at a fine spatial grid over the study area, and at each week of 2009, from data collected over municipalities and months (i.e., we want to simultaneously disaggregate in space and time).

Figure 4.3a shows a fine grid composed of 4871 regular cells of size 1000×1000 m. The blue dots represent the spatial coordinates of the centroids of these cells, and Figure 4.3b shows the spatial distribution of the population on this fine grid (obtained from CBS - Statistics Netherlands, https://www.cbs.nl/nl-nl.), which is heterogeneous across municipalities. To provide a more detailed impression of the Q fever incidence in 2009, we apply the CLMM approach on data recorded over municipalities and months. We choose the fine grid displayed in Figure 4.3a as the spatial refinement, and the weeks of 2009 as the temporal fine scale. To set up the CLMM formulation, we use the spatial coordinates of the grid points as spatial covariates at fine resolution, i.e., $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, the vector $\boldsymbol{x}_3 = (1, ..., 53)'$ as the temporal covariate at fine resolution (since 53 weeks are observed in 2009), second order penalties, 12 equally-spaced knots for the marginal cubic B-spline bases $\mathbf{B}_1$ and $\mathbf{B}_2$ (associated to the spatial covariates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$,

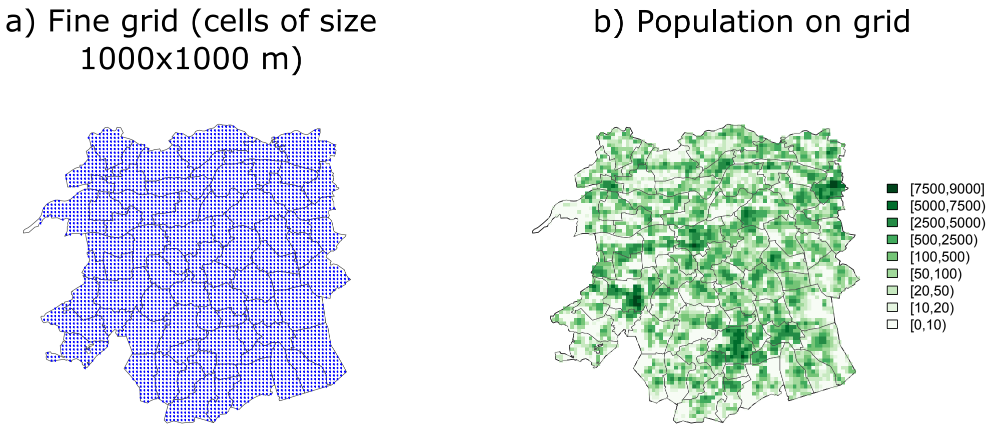


Figure 4.3: The left map shows the fine grid of cell sizes 1000×1000 m in the study area showed in Figure 4.2b. The right map shows the spatial distribution of the population on this fine grid.

respectively), and 8 equally-spaced knots for the marginal cubic B-spline basis $\mathbf{B}_3$ (associated to the temporal covariate $\boldsymbol{x}_3$). We assume here that the population at the fine grid is constant over the time period; thus $\boldsymbol{e}_{\mathrm{f}}$ in Eq. (4.4) is considered as a vector obtained by repeating the fine-scale population fifty three times. The spatial composition matrix is obtained as is described in Eq. (3.5), whereas the temporal composition matrix for this case has the following form:

$$\mathbf{C}_{\mathrm{t}} = \begin{bmatrix}
1 & 1 & 1 & 1 & 1 \\
& & & & & 1 & 1 & 1 & 1 \\
& & & & & & & & & 1 & 1 & 1 & 1 & \frac{3}{7} \\
& & & & & & & & & & & & & \frac{4}{7} & 1 & 1 & 1 & \frac{5}{7} \\
& & & & & & & & & & & & & & & & & \frac{2}{7} & 1 & 1 & 1 & 1 & \frac{1}{7} \\
& & & & & & & & & & & & & & & & & & & & & & \frac{6}{7} & 1 & 1 & 1 & \frac{3}{7} \\
& & & & & & & & & & & & & & & & & & & & & & & & & & \frac{4}{7} & 1 & 1 & 1 & \frac{6}{7} \\
& & & & & & & & & & & & & & & & & & & & & & & & & & & & & & \frac{1}{7} & 1 & 1 & 1 & 1 & \frac{2}{7} \\
& & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & \frac{5}{7} & 1 & 1 & 1 & \frac{4}{7} \\
& & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & \frac{3}{7} & 1 & 1 & 1 & 1 \\
& & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & 1 & 1 & 1 & 1 & \frac{2}{7} \\
& & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & & \frac{5}{7} & 1 & 1 & 1 & 1
\end{bmatrix},$$

where Sunday is considered as the first day of the week. As opposed to the spatial composition matrix, matrix $\mathbf{C}_{\mathrm{t}}$ has some entries that are fractions, this is because some months share parts of a specific week (for example, some days of week 14 belong to March and other to April).

Figure 4.5 shows the resulting CLMM Q fever incidence (per 100000 inhabitants) at the desired fine spatial resolution, for six selected weeks. These incidences

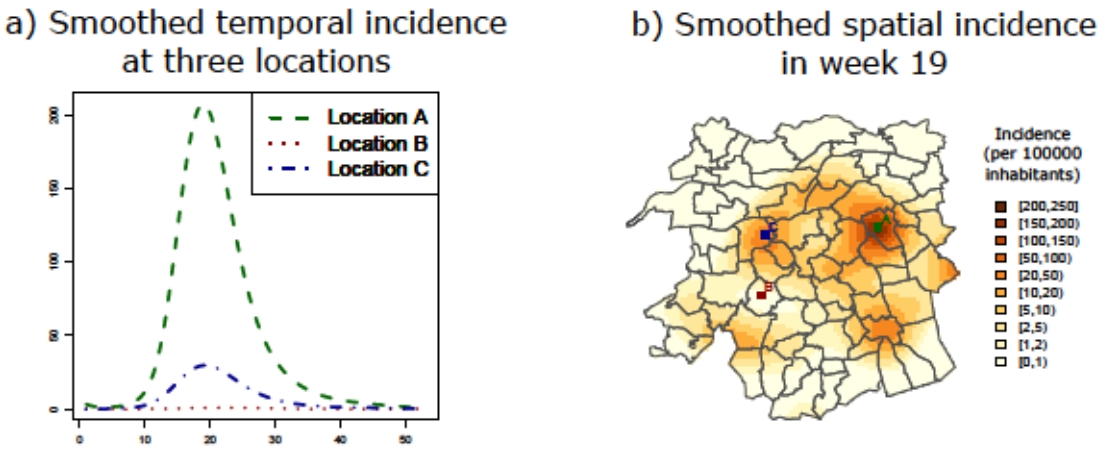


Figure 4.4: The right figure shows the temporal evolution of Q fever incidence in three specific points (A, B, and C), spatially depicted on the map at the left.

are obtained as $\widehat{inc} = 100000 \times \exp(\mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\boldsymbol{\alpha}})$. The evolution of the incidence vary across municipalities and weeks, where higher incidences are mostly observed around week 19. Most of these weeks belong to the months of April, May, and June, which have the largest observed number of Q fever outbreaks in 2009 (see Figure 4.1). Notice also that most of the higher incidences in week 19 are spatially concentrated around the area that involves points A and C in Figure 4.5, which are located in the municipalities of Landerd and Heusden, respectively (see Figure 4.2b). Figure 4.6 shows the approximate standard error maps associated with Figure 4.5. We observe that higher errors are located at the boundary of the study area and at the beginning and the end of the time period. This is because we have less information in these parts, and thus the CLMM estimates are obtained with less precision.

From the previous CLMM estimates, we can also visualize the temporal evolution of the Q fever disease at a specific spatial coordinate of the fine grid. Figure 4.4a shows the smoothed temporal incidence (per week) at three random locations A, B, and C, of the study area. We observe the temporal evolution of the incidence in point B is constant and almost zero, whereas, in points A and C, the temporal smoothed incidence present a unimodal behaviour, where the peak is reached around week 19 (of the month of May). Figure 4.4b shows the spatial trend of the Q fever incidence in this week.

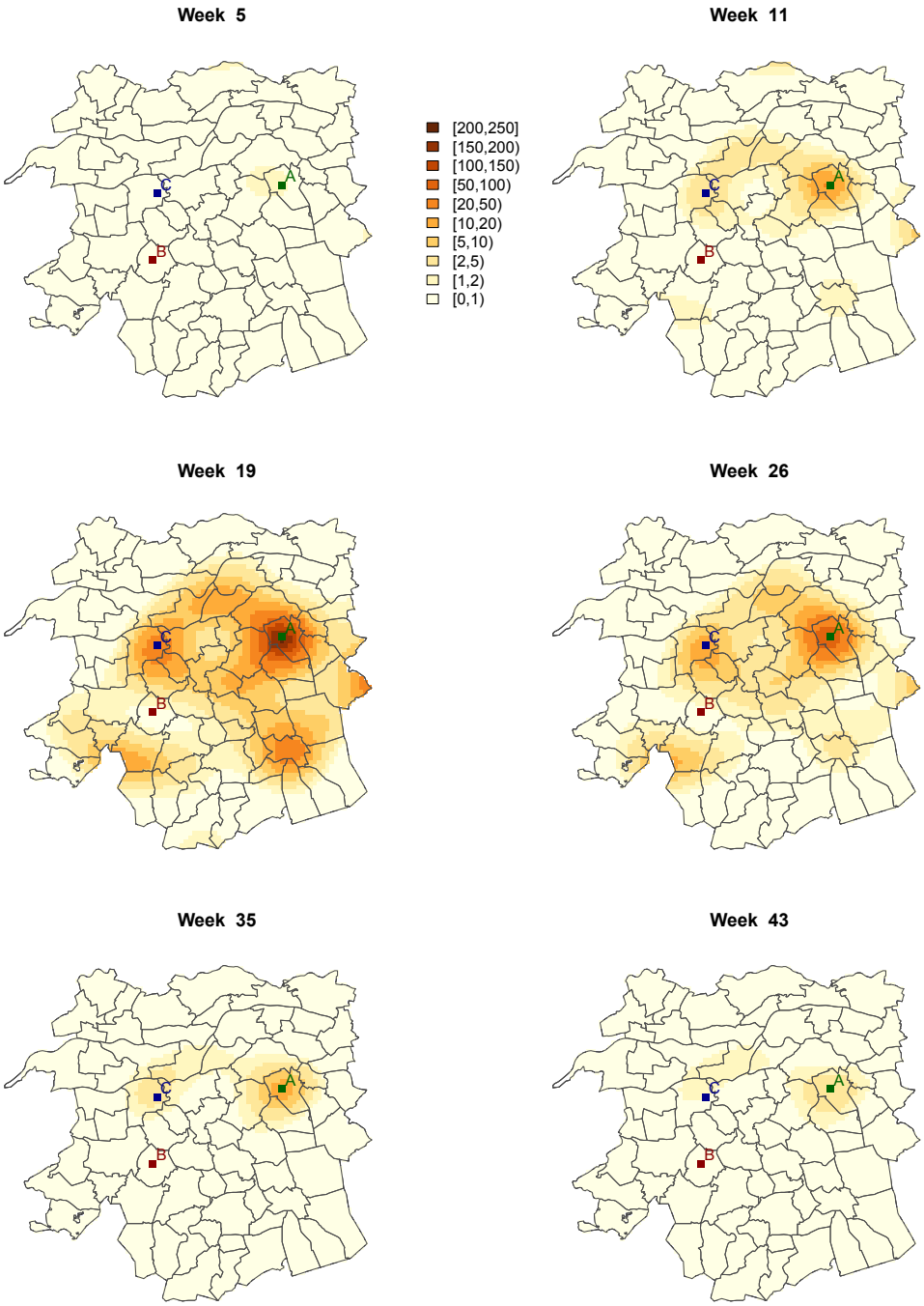


Figure 4.5: Smoothed Q fever incidence at a detailed spatio-temporal scale, resulting from the CLMM approach, for six selected weeks.

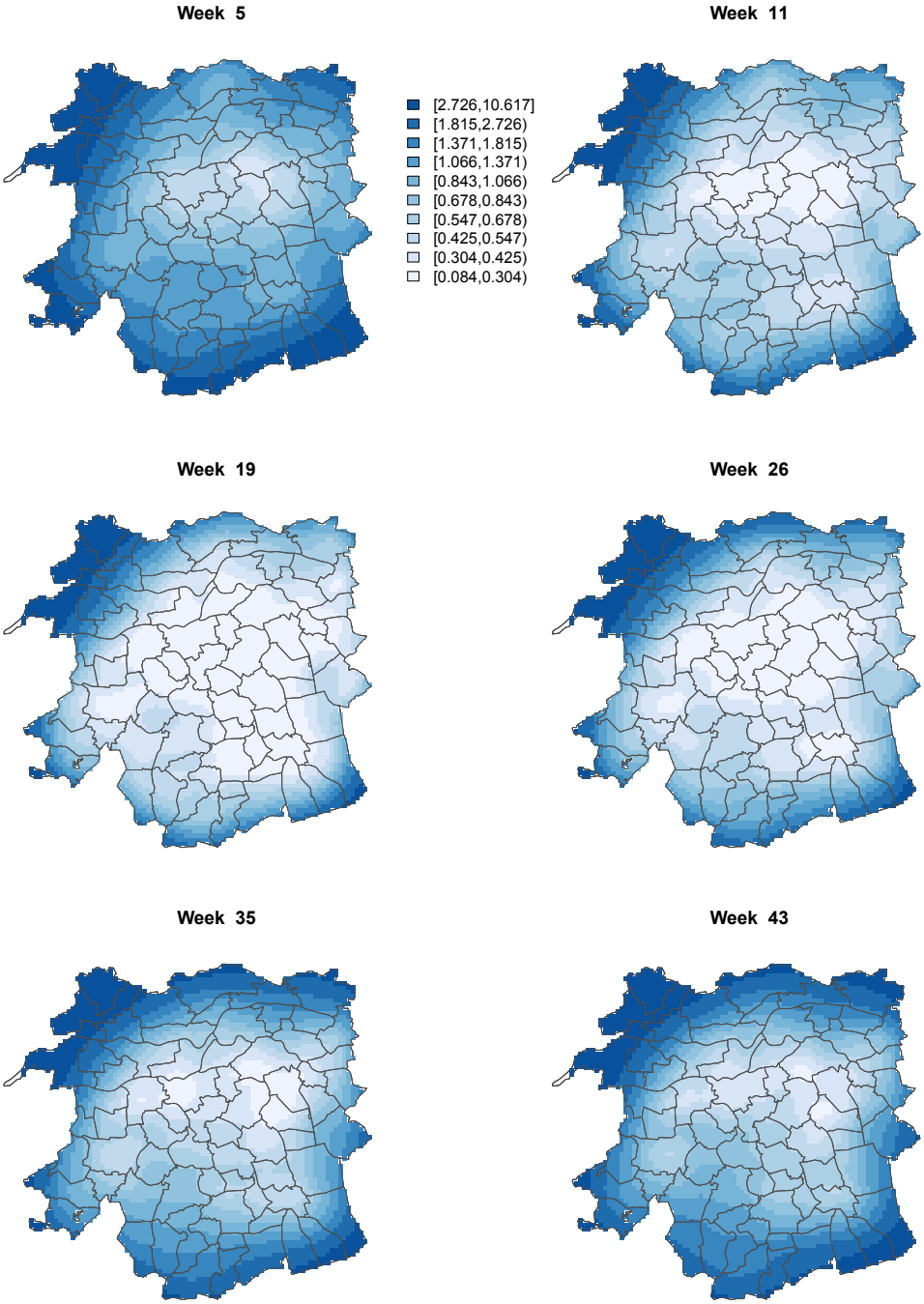


Figure 4.6: Approximate standard error maps associated to the smoothed Q fever incidence maps in Figure 4.5.

4.5 Simulation study

In this section we perform a simulation study in order to examine the prediction performance of the spatio-temporal CLMM approach described in Section 4.1. For that we use a detailed dataset, where the full postal code of the home address and the date of onset of illness of the patients suffering from acute Q fever in the Netherlands, 2009, are available. This dataset is publicly restricted due to confidentiality reasons, and it is only used here to asses how well the CLMM approach recovers the true latent Q fever incidence, when data are available at a specific coarse scale.

Using the full postal code of home address of patients, we can determine the number of cases occurred per week and in each cell of the fine grid depicted in Figure 4.3a. Thus, using the population on the fine grid depicted in Figure 4.3b (and assuming that is constant over weeks), we can obtain a smoothed Q fever incidence surface at this fine spatio-temporal scale using the PGLMM approach, where the spatial coordinates of the fine grid in Figure 4.3a are used as spatial covariates, and the number of weeks is used as temporal covariate. This smoothed incidence surface is considered here as the true latent incidence trend at the fine resolution. We denote these smoothed incidences as $inc(\boldsymbol{u}_j)$, where $\boldsymbol{u}_j$, $j = 1, ..., J$, with $J = 4371 \times 53 = 258163$, represents the spatio-temporal coordinates at fine resolution. To study the prediction performance of the spatio-temporal CLMM approach, we artificially aggregate these Q fever incidence estimates by municipalities and by several coarse periods of time.

The simulation study was conducted as follows:

1. The fine-scale smoothed incidences described above and the population on the fine grid depicted in Figure 4.3b (assumed constant over weeks) were used to calculate the Q fever incidence for each municipality v_i, $i = 1, ..., 72$ (which are depicted in Figure 4.2b), and in different coarse temporal resolutions: 1) fortnights (two weeks); 2) months; and 3) bimesters (two months). Thus we have three types of spatio-temporal aggregation: $g = 1$ where data are summarized over municipalities and fortnights, $g = 2$ where data are summarized over municipalities and months, and $g = 3$ where data are summarized over municipalities and bimesters.

2. 100 realizations of the number of cases recorded over municipalities and each coarse temporal resolution were generated by random drawing of a Poisson distribution whose mean parameter is calculated as the corresponding aggregated incidence (obtained in the previous step) times the population recorded at the appropriate coarse resolution.

3. For each realization, we apply the spatio-temporal CLMM approach using the population on the fine grid (repeated 53 times) as the vector $\boldsymbol{e}_f$ of exposures at the fine resolution.

For all $l = 1, ..., 100$ realizations, the predicted incidence $inc_{P_g}^{(l)}(\boldsymbol{u}_j)$ obtained from the spatio-temporal CLMM approach of each type of aggregation g, with $g = 1, 2, 3$, were compared to the smoothed incidences $inc(\boldsymbol{u}_j)$, $j = 1, ..., J$, using the following criteria:

- Mean absolute error (MAE):

$$\text{MAE}_g^{(l)} = \frac{1}{J}\sum_{j=1}^{J}\left|inc_{P_g}^{(l)}(\boldsymbol{u}_j) - inc(\boldsymbol{u}_j)\right|$$

- Root mean squared error (RMSE):

$$\text{RMSE}_g^{(l)} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\left(inc_{P_g}^{(l)}(\boldsymbol{u}_j) - inc(\boldsymbol{u}_j)\right)^2}$$

Figure 4.7 shows these resulting errors via box-plots for the different types of aggregations, and Table 4.2 gives the averages and the standard deviations of the resulting errors (for each criterion) derived from the simulation study. As we could expect we have found the CLMMs estimates that were obtained from the most coarser spatio-temporal aggregation (type of aggregation 3) are less similar to the true incidences. Notice that these are overall results and the spatial support remains the same in all type of aggregations (municipalities). Similar performances as in Figure 4.7 are obtained if the errors are analysed by weeks. Since, as far as we are aware, no methodologies exists for the simultaneous disaggregation of health data in space and time, we cannot compare the prediction performance of our approach with other techniques.

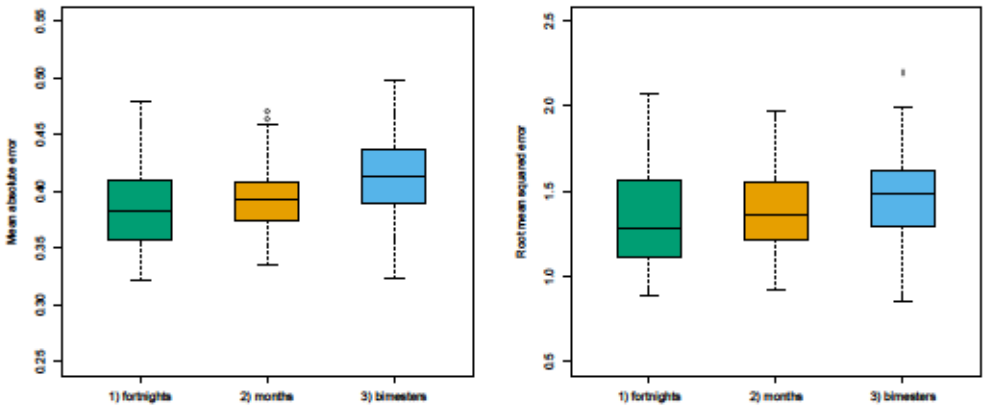


Figure 4.7: Performance comparison of the spatio-temporal CLMM approach under three different types of aggregation (1: municipalities and fortnights; 2: municipalities and months; 3: municipalities and bimesters), using different criteria: mean absolute errors (left) and root mean squared errors (right).

Type of	MAE		RMSE	
aggregation	avg	std	avg	std
$g = 1$: fortnights	0.3856	0.0321	1.3413	0.2774
$g = 2$: months	0.3935	0.0276	1.3933	0.2327
$g = 3$: bimesters	0.4142	0.0326	1.4625	0.2469

Table 4.2: Performance comparison of the spatio-temporal CLMM approach under three different types of aggregation (1: municipalities and fortnights; 2: municipalities and months; 3: municipalities and bimesters), using different criteria: mean absolute errors (MAE) and root mean squared errors (RMSE). These errors are summarized in terms of the average (avg) and standard deviation (std).

4.6 Summary of the chapter

In this chapter we presented and applied the CLMM approach to the disaggregation of grouped data both in space and time. The model allows us to obtain detailed trends of disease incidence, mortality risks, or any other vital rates at a desirable fine spatio-temporal resolution. Thus, the resulting CLMM estimates can be displayed as a dynamic map. Also it allow us to include population infor-

mation at a fine resolution into the estimation process. Here we illustrated the case when this fine scale population is available at a fine regular grid. We performed a simulation study to see the prediction accuracy of the spatio-temporal CLMM using Q fever data recorded at different temporal resolutions (keeping fixed the spatial support). As it was expected, our approach was able to properly capture the underlying trend when the original spatio-temporal resolution is not so coarse.

It is important to acknowledge the use of the SAP algorithm in Section 4.2, together with the GLAM methods (whose usage was described in Section 4.3), to avoid storage problems and to speed up computations. However, we are aware that the disaggregation of grouped data into a very detailed resolution could lead to storage problems and the increase of the computational burden. The sparsity of the marginal composition matrices can be exploited here to deal with these issues.

Chapter 5

Conclusions and further work

Summary of contributions of the thesis

Aggregated data frequently appear in areas such as demography, epidemiology, and public health. For example, it is common to encounter death counts grouped by five-year age classes, followed by an open-ended age class that contains all of the elderly starting at age 80 year or more. In a spatial setting, vital rates are usually collected at a coarse spatial scale formed by irregular geographical units, like counties, districts, and municipalities. Moreover, these rates can be recorded over time, making their analysis more challenging.

In general, data aggregation is done to protect the privacy of patients, to facilitate compact presentation, or to make it comparable with other coarse datasets. This aggregation process may, however, hinder the visualization of the underlying pattern that follow the data. Also, it prohibit the direct analysis of relationships between aggregated data and potential risk factors, which are commonly measured at a finer resolution. Therefore, suitable statistical methodologies are needed in order to estimate the underlying distribution behind aggregated data at a desirable fine scale. In a spatial context, for example, the goal might be to obtain mortality or morbidity estimates, from observed areal data, at a fine spatial grid or at finer areal unit level. These cases are called the area-to-point (ATP) and area-to-area (ATA) cases, respectively, which were illustrated in Chapter 1. To estimate such underlying distribution behind aggregated data, in this thesis we proposed the use of the composite link mixed model (CLMM) approach, which generalizes the penalized composite link model (PCLM) of Eilers (2007) into a

mixed model framework. Our proposal allows to include fine-scale population information (to analyse rates instead of counts) and complex structures as random effects as parts of the modelling of the underlying trend.

Chapter 2 is devoted to present the main aspects of the CLMM approach. First, we briefly reviewed the PCLM of Eilers (2007), which can be seen as a combination of the composite link model approach of Thompson and Baker (1981) and the P-spline methodology of Eilers and Marx (1996). Thus, the flexibility of the model is provided by the use of a B-spline basis as a regression basis, together with a discrete penalty matrix. Then, the CLMM approach is obtained by following the mixed model reformulation of P-splines (proposed by Currie and Durbán, 2002, and Currie et al., 2006), which is based in the use of the singular value decomposition of the penalty matrix. Once the CLMM was presented, we provided a model estimation procedure in Section 2.2, where estimates for the fixed and random effects coefficients of the model were obtained by using the penalized quasi-likelihood (PQL, Breslow and Clayton, 1993) method, and an optimal smoothing parameter value by maximizing the approximate REML given in (2.37). Then, in Section 2.3 the CLMM approach was extended to the multidimensional (array) setting, where the two and three-dimensional cases were illustrated using datasets related with mortality and fertility, respectively.

In Chapter 3 we presented a new methodology for spatially aggregated data, which extends the CLMM approach given in Chapter 2 to the spatial setting. Here, we provided solutions for the ATA and ATP cases previously discussed, and we illustrated them by using several datasets that commonly appear in public health. The spatial CLMM provides a flexible descriptive tool for epidemiological or demographical studies, when the aim is to visualize the spatial distribution of certain rates at a desirable fine resolution. The CLMM approach filters the existing noise in raw rates, which is caused by the small number problem, and allows the creation of more refined mortality or (morbidity) maps by including the distribution of the exposure variable at fine resolution. The resulting CLMM estimates may be linked with potential risks factors that are available over the fine resolution, allowing a posterior correlation analysis between them. Under this framework, we included individual random effects at the aggregated scale to take into account the overdispersion problem, commonly occurring in count data. These individual random effects can be easily included at the fine scale (for graphical representa-

tion) by means of the Moore-Penrose inverse of the composition matrix. Since the CLMM is flexible, no assumptions about the covariance structure of the spatial process should be made (in contrast to kriging methods). The penalty on the coefficients accounts for estimating the spatial trend and the amount of smoothing on each longitude and latitude dimensions. For irregular domains (such as it was the case of the northern Scottish counties and the presence of discontinuities or islands), a possible solution in the CLMM approach is the use of special penalties over complex domains as in Wood et al. (2008), where smoothers are designed to not smooth across boundary features.

We performed a simulation study to compare the ATP Poisson kriging of Goovaerts (2006) with our proposal in Chapter 3, using aggregated data measured over the 92 counties of Indiana and the high-resolution population estimates over a fine grid. The simulation results showed that our proposal is competitive with respect to this geostatistical technique. An additional simulation study using the Scottish lip cancer dataset, where the counties greatly vary in shape and size, is detailed in Appendix B. Here, while the accuracy of the CLMM model is better than the ATP Poisson kriging, further research can be done to improve the smoothing in irregular domains.

In Chapter 4 we generalized the methodology presented in Chapter 3 to the disaggregation of grouped data both in space and time. The resulting spatio-temporal model enables to estimate disease incidence or mortality trends at a fine spatio-temporal resolution, from health data recorded at coarse geographical units and time intervals (for example, from counties and months, to a fine spatial grid and weeks). These estimates, then, can be displayed as a dynamic map, allowing the detailed visualization of the evolution of incidences and mortality trends over time. However, the addition of the temporal dimension in our approach leads to the incrementation of the computational burden (into the model estimation procedure) and storage problems (when we are dealing with a considerable amount of data, or when we want to disaggregate data into very detailed resolution). Thus, in Section 4.3 we provided an alternative procedure for smoothing parameter estimation (into the CLMM framework) based on the SAP algorithm given by Rodríguez-Álvarez et al. (2015). In the adapted SAP algorithm, the smoothing parameters are seen as ratios of variance components and closed-form expressions are derived for them. In combination with the use of adapted GLAM algorithms

presented in Section 4.2, we provided an efficient CLMM estimation procedure under a spatio-temporal context. Moreover, the sparsity of the marginal composition matrices can be exploited here to speed up computations even more (see, for example, Bates and Maechler, 2015). Finally, we performed a simulation study to see the prediction accuracy of the CLMM using Q fever data (which was described in Section 4.4.1) recorded at different temporal resolutions (keeping fixed the spatial support). As it was expected, our approach is able to properly capture the underlying trend when the original spatio-temporal resolution is not so coarse.

Further research

In this thesis we showed the usefulness of the composite link mixed model approach for the estimation of latent trends, from spatially or spatio-temporally aggregated health data. Thus, our approach provides a solution for the *indirect observation* (or *inverse*) *problem* in statistical modelling, in the two and three-dimensional settings. In some cases, however, solutions for the inverse problem to the four-dimensional case are needed. For example, consider mortality data recorded over coarse units and age classes (with different lengths) over time. In this case, a researcher would be interested in to analyse the evolution of the mortality risk on a detailed spatio-temporal scale for each single year old. The composite link mixed model approach can be generalized in order to deal with these complex situations.

In Section 3.2 we presented a methodology to deal with the problem of overdispersion within the (Poisson) composite link mixed model context, where individual random effects have been included at the aggregated scale. Another alternative to deal with that problem would be to develop the composite link mixed model approach under the Negative Binomial framework. The Negative Binomial distribution enables a more flexible modelling of the variance than the Poisson distribution, which can be derived through a Poisson-Gamma mixture (see, for example, Cameron and Trivedi, 1998)

In the area-to-point case, the prediction performance of the composite link mixed model approach for spatially aggregated data was compared with the area-to-point Poisson kriging of Goovaerts (2006). The simulation results showed that our approach is competitive with respect to this geostatistical technique, when the geographical units are similar in shape and size or not. However, further comparisons have to be done, specially with respect to hierarchical Bayesian techniques.

Recently, Taylor et al. (2015) developed an package called , where a continuous log Gaussian Cox process model (Diggle et al., 2013) for areal count data is incorporated within a Bayesian inferential framework. A future goal, then, is to compare our approach with this methodology.

In the illustrations and applications presented in this thesis, we do not have included factor variables such as sex into the analysis. Therefore, it would be of interest to incorporate the factor variable sex in the proposed methodology, in order to simultaneously obtain fine-scale mortality (or morbidity) estimates for each sex. This implies an increment of the number of smoothing parameters to be estimated (if we assume different amounts of smoothness for each sex). The adapted SAP algorithm presented in Chapter 4 will be play a useful role for the smoothing parameter estimation in those cases, specially when we are working with spatially aggregated health data.

In most of the spatio-temporal modelling methodologies existing in the literature, the boundaries of the geographical units are assumed fixed along time. However, this assumption is not always accomplished. For example, when data are collected by postcode, it is common practice to aggregate or disaggregate postcodes over the years depending on the change in population or the urban development; therefore, the geographical units that summarize the data change and in many cases overlap. This issue is known in the literature as the temporal misalignment problem, where some works were proposed to solve it (see, for example, Zhu et al., 2000; Zhu and Carlin, 2000; Hund et al., 2012). As future work we plan to extend the composite link mixed model approach in order to handle this problem, where the marginal composition matrices will play an important role.

From the formulation of the composite link mixed model given in Eq. (2.25), the sum of the latent expectations $\boldsymbol{\gamma}$ is equal to the sum of the aggregated counts $\boldsymbol{y}$. However, the sum of the corresponding latent expectations of each group is not the same as the aggregated count. A solution would be to impose restrictions into the composite link mixed model formulation by means of Lagrange multipliers.

To use the composite link mixed model approach to estimate latent trends from grouped data, we have to know the structure of the composition matrix in advance (which is usually sparse). If its sparse nature is not taken into account (specially in a multidimensional setting), it would lead to an increase in computational time and storage problems. In Chapter 4 we presented efficient algorithms to overcome these

issues under a spatio-temporal context. In order to avoid the explicit construction of the composition matrix, we plan to explore other alternatives such as an adapted version of the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). This idea was explored under a univariate case (see, for example, Uh and Eilers, 2011), but, as far as we are aware, not in the spatial or spatio-temporal case.

References

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons, New Jersey.

Ayma, D., Durbán, M., Lee, D.-J., and Eilers, P. (2016). Penalized composite link models for aggregated spatial count data: A mixed model approach. *Spatial Statistics*, 17:179 – 198.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall, Boca Raton, Florida.

Bates, D. and Maechler, M. (2015). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.2-0.

Bauer, C., Wakefield, J., Rue, H., Self, S., Feng, Z., and Wang, Y. (2016). Bayesian penalized spline models for the analysis of spatio-temporal count data. *Statistics in Medicine*, 35(11):1848–1865.

Berke, O. (2004). Exploratory disease mapping: kriging the spatial risk function from regional count data. *International Journal of Health Geographics*, 3(1):18.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20.

Bindhu, V. M. and Narasimhan, B. (2015). Development of a spatio-temporal disaggregation method (DisNDVI) for generating a time series of fine resolution NDVI images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 101:57–68.

Bivand, R. S. and Lewin-Koh, N. (2016). *maptools: Tools for Reading and Handling Spatial Objects*. R package version 0.8-39.

Bivand, R. S., Pebesma, E. J., and Gómez-Rubio, V. (2008). *Applied Spatial Data Analysis with R*. Springer Science & Business Media.

Braun, J., Duchesne, T., and Stafford, J. E. (2005). Local likelihood density estimation for interval censored data. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 33(1):39–60.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Brewer, C. A., Hatchard, G. W., and Harrower, M. A. (2003). Colorbrewer in print: A catalog of color schemes for maps. *Cartography and Geographic Information Science*, 30(1):5–32.

Brumback, B. A., Ruppert, D., and Wand, M. P. (1999). Comment on: Variable selection and function estimation in additive nonparametric regression using a data-based prior. *Journal of the American Statistical Association*, 94:794–797.

Camarda, C. G., Eilers, P. H. C., and Gampe, J. (2008). Modelling general patterns of digit preference. *Statistical Modelling*, 8(4):385–401.

Camarda, C. G., Eilers, P. H. C., and Gampe, J. (2016). Sums of smooth exponentials to decompose complex series of counts. *Statistical Modelling*.

Cameron, A. C. and Trivedi, P. K. (1998). *Regression Analysis of Count Data*. Econometric Society Monograph No.53. Cambridge University Press, second edition.

Candy, S. G. (1989). Growth and yield models for Pinus Radiata in Tasmania. *New Zealand Journal of Forestry Science*, 19:112–133.

Candy, S. G. (1997). Estimation in forest yield models using composite link functions with random effects. *Biometrics*, 53:146–160.

Caudeville, J., Bonnard, R., Boudet, C., Denys, S., Govaert, G., and Cicolella, A. (2012). Development of a spatial stochastic multimedia exposure model to assess population exposure at regional scale. *Science of The Total Environment*, 432:297–308.

Chilés, J.-P. and Delfiner, P. (1999). *Geostatistics: Modelling Spatial Uncertainty*. Wiley Series in Probablity and Statistics.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43(3):671–681.

Coull, B. A., Schwartz, J., and Wand, M. P. (2001). Respiratory health and air pollution: Additive mixed model analyses. *Biostatistics*, 2:337–349.

Cressie, N. A. C. (1993). *Statistics for Spatial Data (revised edition)*. John Wiley & Sons, New York.

Currie, I. D. and Durbán, M. (2002). Flexible smoothing with P-splines: a unified approach. *Statistical Modelling*, 2(4):333–349.

Currie, I. D., Durbán, M., and Eilers, P. H. C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2):259–280.

de Boor, C. (2001). *A Practical Guide to Splines (revised edition)*. Springer, New York.

de Rooi, J. J., van der Pers, N. M., Hendrikx, R. W. A., Delhez, R., Böttger, A. J., and Eilers, P. H. C. (2014). Smoothing of X-ray diffraction data and Kα_2 elimination using penalized likelihood and the composite link model. *Journal of Applied Crystallography*, 47(3):852–860.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1):1–38.

Dierckx, P. (1993). *Curve and Surface Fitting with Splines*. Clarendon Press, Oxford/Monographs on Numerical Analysis.

Diggle, P. J., Moraga, P., Rowlingson, B., and Taylor, B. M. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563.

Dittrich, R., Francis, B., Hatzinger, R., and Katzenbeisser, W. (2012). Missing observations in paired comparision data. *Statistical Modelling*, 12:117–143.

Domínguez-Berjón, M. F., Borrell, C., Cano-Serral, G., Esnaola, S., Nolasco, A., Pasarín, M. I., Ramis, R., Saurina, C., and Escolar-Pujolar, A. (2008). Construcción de un índice de privación a partir de datos censales en grandes ciudades españolas (Proyecto MEDEA). *Gaceta Sanitaria*, 22:179 – 187.

Eilers, P. (2007). Ill-posed problems with counts, the composite link model and penalized likelihood. *Statistical Modelling*, 7(3):239–254.

Eilers, P. H. C. (2012). Composite link, the neglected model. In *Proceedings of the 27th International Workshop on Statistical Modelling, Prague, Czech Republic*.

Eilers, P. H. C., Currie, I. D., and Durbán, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics & Data Analysis*, 50(1):61–76.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121.

Eilers, P. H. C. and Marx, B. D. (2010). Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:637–653.

Eilers, P. H. C., Marx, B. D., and Durbán, M. (2015). Twenty years of P-splines. *SORT-Statistics and Operations Research Transactions*, 39(2):149–186.

Elliott, P. and Wartenberg, D. (2004). Spatial epidemiology: current approaches and future challenges. *Environmental Health Perspectives*, 112(9):998–1006.

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, 14:731–761.

Faraway, J. J. (2006). *Extending the Linear Models with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall/CRC Texts in Statistical Science.

Fritsch, F. N. and Carlson, R. E. (1980). Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, 17(2):238–246.

Galecki, A. T., Then Have, T. R., and Molenberghs, G. (2001). A simple and fast alternative to the EM algorithm for incomplete categorical data and latent class models. *Computational Statistics & Data Analysis*, 35:265–281.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York.

Goovaerts, P. (2005). Geostatistical analysis of disease data: estimation of cancer mortality risk from empirical frequencies using Poisson kriging. *International Journal of Health Geographics*, 4(1):31.

Goovaerts, P. (2006). Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point Poisson kriging. *International Journal of Health Geographics*, 5:52.

Goovaerts, P. (2008). Kriging and semivariogram deconvolution in presence of irregular geographical units. *Mathematical Geology*, 40(1):101–128.

Gotway, C. A. and Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458):632–648.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review*, 55(3):245–259.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338.

Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective*. Springer-Verlag, New York.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability.

Hund, L., Chen, J. T., Krieger, N., and Coull, B. A. (2012). A geostatistical approach to large-scale disease mapping with temporal misalignment. *Biometrics*, 68(3):849–858.

Jacquez, G. M., Goovaerts, P., Kaufmann, A., and Rommel, R. (2014). *SpaceStat 4.0 User Manual: Software for the Space-Time Analysis of Dynamic Complex Systems*. Publisher: BioMedware, Fourth edition.

Kammann, E. E. and Wand, M. P. (2003). Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52:1–18.

Kelsall, J. and Wakefield, J. (2002). Modeling spatial variation in disease risk: a geostatistical approach. *Journal of the American Statistical Association*, 97(459):692–701.

Koenker, R. and Ng, P. (2016). *SparseM: Sparse Linear Algebra*. R package version 1.72.

Kyriakidis, P. C. (2004). A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis*, 36:259–289.

Lambert, P. (2011). Smooth semiparametric and nonparametric Bayesian estimation of bivariate densities from bivariate histogram data. *Computational Statistics & Data Analysis*, 55(1):429 – 445.

Lambert, P. and Eilers, P. (2009). Bayesian density estimation from grouped continuous data. *Computational Statistics & Data Analysis*, 53(4):1388 – 1399.

Lee, D. J. (2010). *Smoothing mixed models for spatial and spatio-temporal data*. PhD thesis, Department of Statistics, Universidad Carlos III de Madrid, Spain.

Lee, D.-J. and Durbán, M. (2009). Smooth-CAR mixed models for spatial count data. *Computational Statistics & Data Analysis*, 53:2968–2979.

Lee, D.-J. and Durbán, M. (2011). P-spline anova-type interaction models for spatio-temporal smoothing. *Statistical Modelling*, 11:49–69.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):381–400.

MacNab, Y. C. and Dean, C. B. (2001). Autoregressive spatial smoothing and temporal spline smoothing for mapping rates. *Biometrics*, 57:949–956.

MacNab, Y. C. and Dean, C. B. (2002). Spatio-temporal modelling of rates for the construction of disease maps. *Statistics in Medicine*, 21(3):347–358.

Martínez-Beneito, M. A., López-Quilez, A., and Botella-Rocamora, P. (2008). An autoregressive approach to spatio-temporal disease mapping. *Statistics in Medicine*, 27(15):2874–2889.

Maurin, M. and Raoult, D. (1999). Q fever. *Clinical Microbiology Reviews*, 12(4):518–553.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, second edition.

Monestiez, P., Dubroca, L., Bonnin, E., Durbec, J. P., and Guinet, C. (2005). Comparison of model based geostatistical methods in ecology: application to fin whale spatial distribution in Northwestern Mediterranean Sea. In Leuangthong, O. and Deutsch, C. V., editors, *Geostatistics Banff 2004*, volume 14 of *Quantitative Geology and Geostatistics*, pages 777–786. Springer Netherlands.

Monestiez, P., Dubroca, L., Bonnin, E., Durbec, J. P., and Guinet, C. (2006). Geostatistical modelling of spatial distribution of *Balaenoptera physalus* in the Northwestern Mediterranean Sea from sparse count data and heterogeneous observation efforts. *Ecological Modelling*, 193:615–628.

Mugglin, A. S. and Carlin, B. P. (1998). Hierarchical modeling in geographic information systems: population interpolation over incompatible zones. *Journal of Agricultural, Biological, and Environmental Statistics*, 3(2):111–130.

Müller, H.-G., Stadtmüller, U., and Tabnak, F. (1997). Spatial smoothing of geographically aggregated data, with application to the construction of incidence maps. *Journal of the American Statistical Association*, 92:61–71.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2.

Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G., and Breidt, F. J. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70:265–286.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545–554.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press.

Perperoglou, A. and Eilers, P. H. C. (2010). Penalized regression with individual deviance effects. *Computational Statistics*, 25(2):341–361.

Pickle, L., Mungiole, M., Jones, G. K., and White, A. A. (1999). Exploring spatial patterns of mortality: the new Atlas of United States Mortality. *Statistics in Medicine*, 18(23):3211–3220.

Pinheiro, J. C. and Bates, D. (2000). *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. Springer, New York.

Prairie, J., Rajagopalan, B., Lall, U., and Fulp, T. (2007). A stochastic nonparametric technique for space-time disaggregation of streamflows. *Water Resources Research*, 43(3).

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabe-Hesketh, S. and Skrondal, A. (2007). Multilevel and latent variable modeling with composite links and exploded likelihoods. *Psychometrika*, 72(2):123–140.

Ramsay, T. (2002). Spline smoothing over difficult regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64:307–319.

Rey, G., Jougla, E., Fouillet, A., and Hémon, D. (2009). Ecological association between a deprivation index and mortality in France over the period 1997-2001: variations with spatial scale, degree of urbanicity, age, gender and cause of death. *BMC Public Health*, 9(33).

Rindskopf, D. (1992). A general approach to categorical data analysis with missing data, using generalized linear models with composite link. *Psychometrika*, 57:29–42.

Rizzi, S., Gampe, J., and Eilers, P. H. C. (2015). Efficient estimation of smooth distributions from coarsely grouped data. *American Journal of Epidemiology*.

Rizzi, S., Thinggaard, M., Engholm, G., Christensen, N., Johannesen, T. B., Vaupel, J. W., and Lindahl-Jacobsen, R. (2016). Comparison of non-parametric methods for ungrouping coarsely aggregated data. *BMC Medical Research Methodology*, 16(1):1–12.

Rodríguez-Álvarez, M. X., Lee, D.-J., Kneib, T., Durbán, M., and Eilers, P. H. C. (2015). Fast smoothing parameter separation in multidimensional generalized P-splines: the SAP algorithm. *Statistics and Computing*, 25(5):941–957.

Roest, H. I. J., Tilburg, J. J. H. C., van der Hoek, W., Vellema, P., van Zijderveld, F. G., Klaassen, C. H. W., and Raoult, D. (2011). The Q fever epidemic in The Netherlands: history, onset, response and reflection. *Epidemiology & Infection*, 139(1):1–12.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Salmond, C. E. and Crampton, P. (2012). Development of New Zealand's deprivation index (NZDep) and its uptake as a national policy tool. *Canadian Journal of Public Health*, 103(Suppl 2):S7–S11.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78(4):719–721.

Schleiss, M. and Berne, A. (2012). Stochastic space-time disaggregation of rainfall into DSD fields. *Journal of Hydrometeorology*, 13(6):1954–1969.

Schrödle, B. and Held, L. (2011). Spatio-temporal disease mapping using INLA. *Environmetrics*, 22(6):725–734.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance components*. Wiley Series in Probability and Mathematical Statistics.

Segond, M.-L., Neokleous, N., Makropoulos, C., Onof, C., and Maksimovic, C. (2007). Simulation and spatio-temporal disaggregation of multi-site rainfall data for urban drainage applications. *Hydrological Sciences Journal*, 52(5):917–935.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer Series in Statistics. Springer-Verlag, New York.

Smith, L., Hyndman, R. J., and Wood, S. N. (2004). Spline interpolation for demographic variables: The monotonicity problem. *Journal of Population Research*, 21(1):95–98.

Taylor, B. M., Davies, T. M., Rowlingson, B. S., and Diggle, P. J. (2015). Bayesian inference and data augmentation schemes for spatial, spatiotemporal and multivariate log-gaussian cox processes in R. *Journal of Statistical Software*, 63(7):1–48.

Thompson, R. and Baker, R. J. (1981). Composite link functions in generalized linear models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 30(2):125–131.

Ugarte, M. D., Adin, A., Goicoa, T., and Militino, A. F. (2014). On fitting spatio-temporal disease mapping models using approximate bayesian inference. *Statistical Methods in Medical Research*, 23(6):507–530.

Ugarte, M. D., Goicoa, T., and Militino, A. F. (2010). Spatio-temporal modeling of mortality risks using penalized splines. *Environmetrics*, 21(3–4):270–289.

Uh, H.-W. and Eilers, P. H. C. (2011). Haplotype estimation from fuzzy genotypes using penalized likelihood. *PLoS ONE*, 6(9).

van den Hout, A., Gilchrist, R., and van der Heijden, P. G. M. (2010). The randomized response log linear model as a composite link model. *Statistical Modelling*, 10:57–67.

van der Hoek, W., Dijkstra, F., Schimmer, B., Schneeberger, P. M., Vellema, P., Wijkmans, C., ter Schegget, R., Hackert, V., and van Duynhoven, Y. (2010). Q fever in the Netherlands: an update on the epidemiology and control measures. *Eurosurveillance*, 15(12):pii:19520.

Ver Hoef, J. (2012). Who invented the Delta method? *The American Statistician*, 66:124–127.

Waller, L. A., Carlin, B. P., Xia, H., and Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, 92(438):607–617.

Waller, L. A. and Gotway, C. A. (2004). *Applied Spatial Statistics for Public Health Data*. John Wiley & Sons, New York.

Wand, M. P. (2003). Smoothing and mixed models. *Computational Statistics*, 18:223–249.

Wang, B. and Wertelecki, W. (2013). Density estimation for data with rounding errors. *Computational Statistics & Data Analysis*, 65:4 – 12.

Wikle, C. K., Berliner, L. M., and Cressie, N. (1998). Hierarchical Bayesian space-time models. *Journal of Environmental and Ecological Statistics*, 5:117–154.

Wood, S. N. (2006a). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC Texts in Statistical Science.

Wood, S. N. (2006b). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, 62:1025–1036.

Wood, S. N., Bravington, M. V., and Hedley, S. L. (2008). Soap film smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70:931–955.

Yavuz, A. C. and Lambert, P. (2011). Smooth estimation of survival functions and hazard ratios from interval-censored data using Bayesian penalized B-splines. *Statistics in Medicine*, 30(1):75–90.

Zhu, L. and Carlin, B. P. (2000). Comparing hierarchical models for spatio-temporally misaligned data using the deviance information criterion. *Statistics in Medicine*, 19(17-18):2265–2278.

Zhu, L., Carlin, B. P., English, P., and Scalf, R. (2000). Hierarchical modelling of spatio-temporally misaligned data: relating traffic density to paediatric asthma hospitalizations. *Environmetrics*, 11(1):43–61.

Appendix A

Appendix to Chapter 2

In this section we present some notation and definitions of array methods proposed in Currie et al. (2006) and Eilers et al. (2006), which we have introduced in Section 2.3.

Definition A.1. The row tensor of a matrix $\mathbf{X}$ with c columns is defined as:

$$\mathcal{G}(\mathbf{X}) = (\mathbf{X} \otimes \mathbf{1}_c') \odot (\mathbf{1}_c' \otimes \mathbf{X}),$$

where $\mathbf{1}_c$ denotes a vector of 1's of length c, and $\odot$ is the element-by-element product.

The previous definition can be extended in the following way.

Definition A.2. The row tensor of the matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, of dimensions $n \times c_1$ and $n \times c_2$, respectively, is defined as:

$$\mathcal{G}(\mathbf{X}_1, \mathbf{X}_2) = (\mathbf{X}_1 \otimes \mathbf{1}_{c_2}') \odot (\mathbf{1}_{c_1}' \otimes \mathbf{X}_2).$$

Note that the previous definition denotes the 'row-wise' Kronecker product of two matrices, which we have introduced in Section 3.1.

Definition A.3. The $\mathcal{H}$-transform of the d-dimensional array $\mathbf{A}$ of size $c_1 \times c_2 \times \cdots \times c_d$ by the matrix $\mathbf{X}$ of dimension $r \times c_1$, denoted as $\mathcal{H}(\mathbf{X}, \mathbf{A})$, is defined as follows. Let $\mathbf{A}^*$ be the matrix of dimension $c_1 \times c_2 c_3 \cdots c_d$ that is obtained by flattening dimensions 2-d of $\mathbf{A}$; form the matrix product $\mathbf{X}\mathbf{A}^*$ of dimension

$r \times c_2 c_3 \cdots c_d$; then $\mathcal{H}(\mathbf{X}, \mathbf{A})$ is the d-dimensional array of size $r \times c_2 \times \cdots \times c_d$ that is obtained from $\mathbf{X}\mathbf{A}^*$ by reinstating dimensions 2-d of $\mathbf{A}$.

If $\mathbf{A}$ is a vector, i.e., $\mathbf{A} = \boldsymbol{a}$, then we have that $\mathcal{H}(\mathbf{X}, \mathbf{A}) = \mathbf{X}\boldsymbol{a}$, whereas if $\mathbf{A}$ is a matrix, $\mathcal{H}(\mathbf{X}, \mathbf{A}) = \mathbf{X}\mathbf{A}$. Thus, the $\mathcal{H}$-transform generalizes premultiplication of vector and matrices by a matrix. The following definition generalizes the transpose of a matrix.

Definition A.4. The rotation of the d-dimensional array $\mathbf{A}$ of size $c_1 \times c_2 \times \cdots \times c_d$ is the d-dimensional array $\mathcal{R}(\mathbf{A})$ of size $c_2 \times c_3 \times \cdots \times c_d \times c_1$ that is obtained by permuting the indices of $\mathbf{A}$.

Combining the last two definitions, we obtain:

Definition A.5. The rotated $\mathcal{H}$-transform of the array $\mathbf{A}$ by the matrix $\mathbf{X}$ is given by:

$$\rho(\mathbf{X}, \mathbf{A}) = \mathcal{R}(\mathcal{H}(\mathbf{X}, \mathbf{A})).$$

Appendix B

Appendix to Chapter 3

In this appendix we include an additional simulation study to compare the prediction performance among CLMM, CLMM-P and PK, when the geographical units vary considerably in shape and size. For that, we use the Scottish lip cancer dataset described in Section 3.3.3. Here we use the estimated vector of naive exposures as the true exposures at fine grid (that is, $\boldsymbol{e}_{\mathrm{f}} = \hat{\boldsymbol{e}}_{\mathrm{naive}}$).

The simulation study was conducted in a similar fashion as in Section 3.3.4, where the continuous mortality risk surface obtained with the PK approach was considered here as the true underlying mortality trend (see Figure 3.4f). Thus, for the resulting 100 realizations, the predicted risks $r_{\mathrm{P}}^{(l)}(\boldsymbol{u}_j)$ obtained from the three approaches were compared to the true underlying mortality risk, using the ME, MAE and RMSE criteria. Figure B.1 shows these resulting errors via box-plots, in which we observe the CLMM and CLMM-P approaches give better prediction accuracy than PK, for each criterion. Note that, in this simulation setting, we did not include any overdispersion, and hence both CLMM and CLMM-P approaches are very similar. Table B.1 gives the averages and the standard deviations of the resulting errors (for each criterion) computed from this additional simulation study.

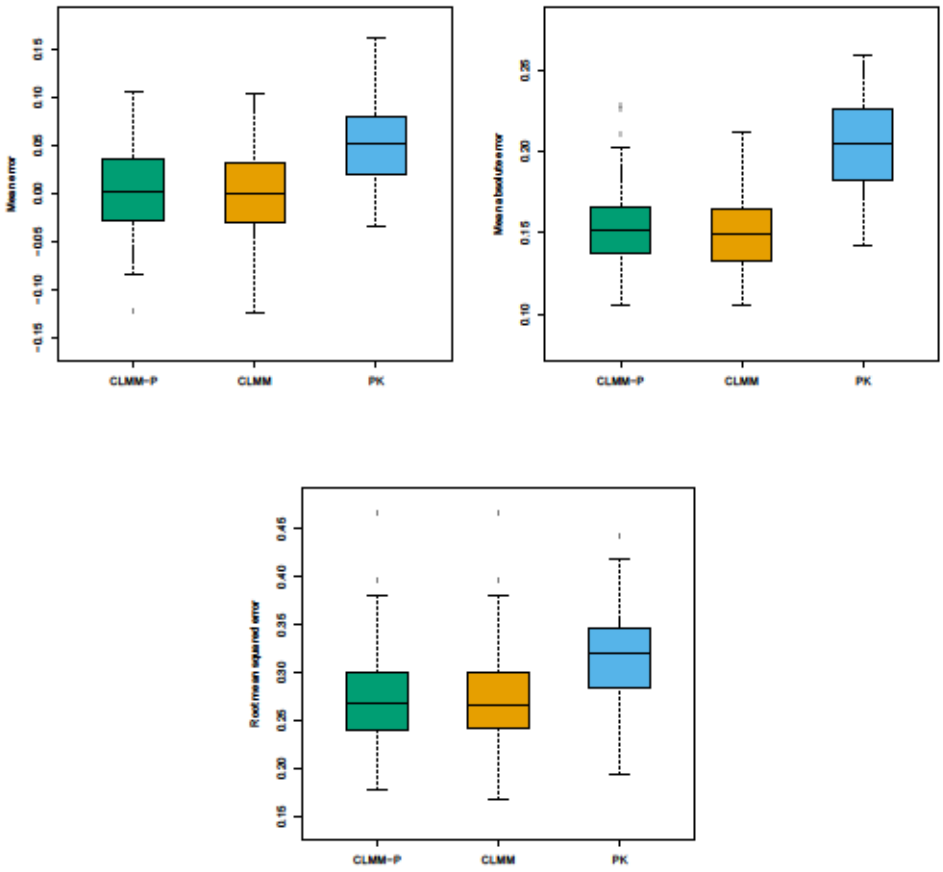


Figure B.1: Performance comparison between CLMM-P, CLMM and PK approaches using different criteria: mean errors (top-left), mean absolute errors (top-right), and root mean squared errors (bottom).

Approach	ME		MAE		RMSE	
	avg	std	avg	std	avg	std
CLMM-P	0.0040	0.0464	0.1523	0.0232	0.2748	0.0512
CLMM	0.0012	0.0463	0.1493	0.0216	0.2749	0.0505
PK	0.0552	0.0423	0.2041	0.0277	0.3191	0.0460

Table B.1: Performance comparison of CLMM-P, CLMM and PK approaches, using different criteria: mean errors (ME), mean absolute errors (MAE), and root mean squared errors (RMSE). These errors are summarized in terms of the average (avg) and standard deviation (std).

Appendix C

Appendix to Chapter 4

In this section we provide the proof for the closed-form expressions of the variance components given in Eq. (4.6). We should note that the following proof is similar to that given in Rodríguez-Álvarez et al. (2015), but now the 'working' mixed model matrices $\breve{\mathbf{X}}$ and $\breve{\mathbf{Z}}$ (as well as related matrices like $\mathbf{V}$ and $\mathbf{N}$) appear. Here the inverse of the covariance matrix $\mathbf{G}$ given in (2.53) with $\lambda_d = \frac{1}{\tau_d^2}$ $(d = 1, 2, 3)$ is used.

Consider the approximate REML version of Patterson and Thompson (1971) (which is equivalent to (2.37)):

$$l^*_{\text{REML}} = -\frac{1}{2}\underbrace{\log|\mathbf{V}|}_{\textbf{Part I}} - \frac{1}{2}\underbrace{\log|\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}}|}_{\textbf{Part II}} - \frac{1}{2}\underbrace{(\boldsymbol{z} - \breve{\mathbf{X}}\widehat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\boldsymbol{z} - \breve{\mathbf{X}}\widehat{\boldsymbol{\beta}})}_{\textbf{Part III}}, \qquad \text{(C.1)}$$

in which the dependence of the matrix of weights $\mathbf{W}$ on τ_d^2 $(d = 1, 2, 3)$ is ignored. To obtain the REML estimates of the variance components τ_d^2, we first differentiate each part of Eq. (C.1) with respect to them. Using the fact that $\frac{\partial \mathbf{V}}{\partial \tau_d^2} = \breve{\mathbf{Z}}\frac{\partial \mathbf{G}}{\partial \tau_d^2}\breve{\mathbf{Z}}'$, with $\mathbf{V}$ defined in Eq. (2.35), we obtain:

Part I: Here we use property (8.6) given in Harville, 1997, p.305:

$$\frac{\partial \log|\mathbf{V}|}{\partial \tau_d^2} = \text{trace}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \tau_d^2}\right) = \text{trace}\left(\mathbf{V}^{-1}\breve{\mathbf{Z}}\frac{\partial \mathbf{G}}{\partial \tau_d^2}\breve{\mathbf{Z}}'\right)$$

Part II: Here we use properties (8.6) and (8.18) given in Harville, 1997, pp. 305,

308:

$$\begin{aligned}\frac{\partial \log|\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}}|}{\partial\tau_d^2} &= \text{trace}\left((\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})^{-1}\frac{\partial(\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})}{\partial\tau_d^2}\right)\\ &= -\text{trace}\left((\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})^{-1}\breve{\mathbf{X}}'\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\tau_d^2}\mathbf{V}^{-1}\breve{\mathbf{X}}\right)\\ &= -\text{trace}\left(\mathbf{V}^{-1}\breve{\mathbf{X}}(\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})^{-1}\breve{\mathbf{X}}'\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\tau_d^2}\right)\\ &= -\text{trace}\left(\mathbf{V}^{-1}\breve{\mathbf{X}}(\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})^{-1}\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{Z}}\frac{\partial\mathbf{G}}{\partial\tau_d^2}\breve{\mathbf{Z}}'\right)\end{aligned}$$

Part III: Here we use property (8.6) given in Harville, 1997, p. 308:

$$\begin{aligned}\frac{\partial(\boldsymbol{z}-\breve{\mathbf{X}}\widehat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\boldsymbol{z}-\breve{\mathbf{X}}\widehat{\boldsymbol{\beta}})}{\partial\tau_d^2} &= -(\boldsymbol{z}-\breve{\mathbf{X}}\widehat{\boldsymbol{\beta}})'\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\tau_d^2}\mathbf{V}^{-1}(\boldsymbol{z}-\breve{\mathbf{X}}\widehat{\boldsymbol{\beta}})\\ &= -(\boldsymbol{z}-\breve{\mathbf{X}}\widehat{\boldsymbol{\beta}})'\mathbf{V}^{-1}\breve{\mathbf{Z}}\frac{\partial\mathbf{G}}{\partial\tau_d^2}\breve{\mathbf{Z}}'\mathbf{V}^{-1}(\boldsymbol{z}-\breve{\mathbf{X}}\widehat{\boldsymbol{\beta}})\\ &= -\boldsymbol{b}\frac{\partial\mathbf{G}}{\partial\tau_d^2}\boldsymbol{b}\\ &= -\boldsymbol{\alpha}'\mathbf{G}^{-1}\frac{\partial\mathbf{G}}{\partial\tau_d^2}\mathbf{G}^{-1}\boldsymbol{\alpha}\end{aligned}$$

Adding the derivatives obtained from Part I and Part II, we obtain:

$$\begin{aligned}\frac{\partial\log|\mathbf{V}|}{\partial\tau_d^2}+\frac{\partial\log|\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}}|}{\partial\tau_d^2} &= \text{trace}\left((\mathbf{V}^{-1}-\mathbf{V}^{-1}\breve{\mathbf{X}}(\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})^{-1}\breve{\mathbf{X}}'\mathbf{V}^{-1})\breve{\mathbf{Z}}\frac{\partial\mathbf{G}}{\partial\tau_d^2}\breve{\mathbf{Z}}'\right)\\ &= \text{trace}\left(\mathbf{N}\breve{\mathbf{Z}}\frac{\partial\mathbf{G}}{\partial\tau_d^2}\breve{\mathbf{Z}}'\right)\\ &= \text{trace}\left(\breve{\mathbf{Z}}'\mathbf{N}\breve{\mathbf{Z}}\frac{\partial\mathbf{G}}{\partial\tau_d^2}\right),\end{aligned}$$

It follows that:

$$\frac{\partial l^*_{\text{REML}}}{\partial\tau_d^2} = -\frac{1}{2}\text{trace}\left(\breve{\mathbf{Z}}'\mathbf{N}\breve{\mathbf{Z}}\frac{\partial\mathbf{G}}{\partial\tau_d^2}\right)+\frac{1}{2}\boldsymbol{\alpha}'\mathbf{G}^{-1}\frac{\partial\mathbf{G}}{\partial\tau_d^2}\mathbf{G}^{-1}\boldsymbol{\alpha}, \tag{C.2}$$

where:

$$\frac{\partial\mathbf{G}}{\partial\tau_d^2} = -\mathbf{G}\frac{\partial\mathbf{G}^{-1}}{\partial\tau_d^2}\mathbf{G} = -\mathbf{G}\left(-\frac{1}{\tau_d^4}\boldsymbol{\Lambda}_d\right)\mathbf{G} = \frac{1}{\tau_d^4}\mathbf{G}\boldsymbol{\Lambda}_d\mathbf{G}, \tag{C.3}$$

for $d = 1, 2, 3$.

By replacing Eq. (C.3) in Eq. (C.2), we obtain:

$$\frac{\partial l^{*}_{\text{REML}}}{\partial \tau^2_d} = -\frac{1}{2\tau^2_d}\text{trace}\left(\breve{\mathbf{Z}}'\mathbf{N}\breve{\mathbf{Z}}\mathbf{G}\frac{\mathbf{\Lambda}_d}{\tau^2_d}\mathbf{G}\right) + \frac{1}{2\tau^4_d}\boldsymbol{\alpha}'\mathbf{\Lambda}_d\boldsymbol{\alpha}.$$

Therefore, the REML estimates of the variance components τ^2_d are found by equating the expression above by zero. These estimates are given by:

$$\hat{\tau}^2_d = \frac{\boldsymbol{\alpha}'\mathbf{\Lambda}_d\boldsymbol{\alpha}}{\text{trace}\left(\breve{\mathbf{Z}}'\mathbf{N}\breve{\mathbf{Z}}\mathbf{G}\frac{\mathbf{\Lambda}_d}{\tau^2_d}\mathbf{G}\right)}, \text{ for } d = 1, 2, 3.$$

Notice that if we add the denominators of the previous expressions, we obtain:

$$\begin{aligned}\sum_{d=1}^{3}\text{trace}\left(\breve{\mathbf{Z}}'\mathbf{N}\breve{\mathbf{Z}}\mathbf{G}\frac{\mathbf{\Lambda}_d}{\tau^2_d}\mathbf{G}\right) &= \text{trace}\left(\sum_{d=1}^{3}\breve{\mathbf{Z}}'\mathbf{N}\breve{\mathbf{Z}}\mathbf{G}\frac{\mathbf{\Lambda}_d}{\tau^2_d}\mathbf{G}\right)\\ &= \text{trace}\left(\breve{\mathbf{Z}}'\mathbf{N}\breve{\mathbf{Z}}\mathbf{G}\right)\\ &= \text{trace}\left(\breve{\mathbf{Z}}\mathbf{G}\breve{\mathbf{Z}}'\mathbf{N}\right),\end{aligned}$$

where $\breve{\mathbf{Z}}\mathbf{G}\breve{\mathbf{Z}}'\mathbf{N}$ is the hat matrix (Hastie and Tibshirani, 1990) of the unpenalized (or random) part of the fitted CLMM (see Eq. (2.33)).